6CCM331A Special Relativity and Electromagnetism Lecture Notes

Dr. Paul P. Cook
Department of Mathematics, King's College London.

27th April 2026

"From a long view of the history of mankind - seen from, say, ten thousand years from now - there can be little doubt that the most significant event of the 19^{th} century will be judged as Maxwell's discovery of the laws of electrodynamics. The American Civil War will pale into provincial insignificance in comparison with this important scientific event of the same decade."

Richard Feynman, The Feynman Lectures on Physics Volume II, 1964

0. Contents

1 Introduction **7**

1.1 Historic introduction 7

1.2 Maths and Physics 8

1.3 Course Organisation 9

1.3.1 Reading List 9

2 A Review of Vector Calculus **11**

2.1 Vector Derivatives. 11

2.1.1 The Gradient. 11

2.1.2 The Divergence. 12

2.1.3 The Curl. 16

2.1.4 Product Rules. 20

2.1.5 Second Derivatives. 23

2.2 Vector Integrals. 24

2.2.1 The Line Integral. 24

2.2.2 The Surface Integral. 29

2.2.3 The Volume Integral. 33

2.2.4 The Fundamental Theorems of Vector Calculus 34

2.2.5 The Dirac Delta Function 38

3 Electrostatics **43**

3.1 Coulomb's Law. 43

3.2 The Electric Field 46

3.2.1 Electric Field Lines 47

3.3 Gauss's Law 50

3.3.1 Application of Gauss's Law 52

3.3.2 Changes in the Electric Field Across a Charged Surface 56

3.4 The Electric Potential 57

3.5 Work and Energy . . . 63
3.6 Conductors . . . 65

4 Potentials **71**
4.1 Poisson and Laplace's Equations . . . 71
4.1.1 Harmonic Functions . . . 72
4.2 Boundary Conditions . . . 73
4.2.1 The Uniqueness of Solutions to Laplace's Equation . . . 74
4.2.2 The Method of Images . . . 76
4.3 Separation of Variables in Laplace's Equation . . . 79
4.4 Poisson's Equation and Green's Functions . . . 85
4.5 The Multipole Expansion . . . 89

5 Magnetostatics **97**
5.1 The Lorentz Force Law . . . 97
5.2 Currents . . . 99
5.3 The Biot-Savart Law and Ampère's Law . . . 101
5.3.1 The Divergence and Curl of $\boldsymbol{B}$ . . . 105
5.4 The Magnetic Vector Potential . . . 112
5.5 The Multipole Expansion for the Vector Potential . . . 115

6 Electrodynamics **117**
6.1 Maxwell's Equations . . . 117
6.1.1 Electromotive Forces and Faraday's Law . . . 117
6.1.2 Correcting Ampère's Law . . . 122
6.2 Electromagnetic Waves in Vacuum . . . 127
6.2.1 Travelling Waves . . . 128
6.2.2 Plane Wave Solutions of Maxwell's Equations . . . 130
6.3 Electromagnetic Energy . . . 136
6.4 Potentials . . . 140
6.5 The Retarded and the Advanced Potentials . . . 142

7 Special Relativity **149**
7.1 The Postulates of Special Relativity . . . 150
7.1.1 Simultaneity . . . 152
7.1.2 Time Dilation and Length Contraction . . . 160
7.2 The Lorentz Transformations . . . 165

7.2.1 Addition of Velocity 167
7.3 Minkowski Space 169
7.4 The Lorentz Group 173
7.4.1 Lorentz Vectors and Tensors 175
7.5 4-Displacement, 4-Velocity and 4-Momentum 177

8 Relativistic Electrodynamics 181
8.1 Lorentz Boosting the Magnetic Field of a Wire 181
8.2 The Lorentz Transformations of $\boldsymbol{E}$ and $\boldsymbol{B}$ 185
8.3 A Lorentz Tensor for Electromagnetism 191
8.4 Maxwell's Equations in Tensor Notation. 195

Appendices

A Index Notation and the Einstein Summation Convention 201

B Scalars and Vectors 211

C Cartesian, Cylindrical and Spherical Polar Coordinates 215
C.1 Cylindrical Coordinates 215
C.2 Spherical Coordinates 220

D Orthogonal Polynomials 223
D.1 Classical Orthogonal Polynomials 224
D.2 The Legendre Polynomials 228

E Tutorial Exercises 231

Acknowledgements

These lecture notes are based on the lecture courses by David Griffiths, Ashok Das, Richard Feynman and David Morin whose books are given in the reading list in the introduction. Nice ideas, cute problems and interesting thoughts are all due to these four authors. Stop reading these notes here and go and read theirs instead. I am grateful to Nadav Drukker, Nikolay Gromov, and Andreas Recknagel for sharing with me their lectures and problems on these topics. I am indebted to Keith Glennon, Saad Altaf, Bechir Houas, Christian Lai, Tobias Dear, Alex Jones, James Carden, Joseph Sullivan, Pooja Amirthalingam, Mengyue Liu, Alisher Mirmanov, Arthur Costa and Mohammad Yaseen Qayyum for pointing out typographical errors and mistakes in earlier versions of these lecture notes. All typographical errors and errors of any other type are, of course, my fault and I will be grateful for all comments on any kind of mistake.

Paul P. Cook, 22$^{\text{nd}}$ November, 2021.

1. Introduction

1.1 Historic introduction

The theory of electromagnetism was discovered by James Clerk Maxwell at King's College London in 1865. Prior to Maxwell's work, the electric force, the magnetic force and all kinds of electromagnetic radiation (light and radio waves for example) were understood to be distinct aspects of the natural world. Maxwell unified electrostatics, the forces due to a static array of electric charges, and magnetostatics, the forces due to a static magnetic field. His unification was brilliant and he did it by understanding that the motion of electric charges interacted with magnetic fields - by introducing dynamics and relative motion into the study of electrostatics and magnetostatics, Maxwell could show that electricity and magnetism are two facets of a single underlying idea: electromagnetism. As a result of his unification, Maxwell discovered that light and radio waves are forms of electromagnetic radiation. Much of the modern world is underpinned by the revolution in thought and engineering that Maxwell's electromagnetism brought about.

Another consequence of Maxwell's work was the discovery that the speed of light is a constant. This idea had to wait half a century before Einstein outlined all the consequences this entailed. In 1905, Albert Einstein published a paper outlining his theory of special relativity. By assuming both that the speed of light is a constant, and the principle of equivalence, Einstein showed that Newtonian laws of motion required corrections. Special Relativity also entailed a number of spectacular predictions, that time runs slower in a frame of reference at motion with respect to your own frame, and similarly that lengths contract in a similar situation. All the predictions of special relativity have been experimentally verified. One consequence of special relativity is the idea that the natural world that humans experience is more simply understood if it is a four-dimensional space-time, rather than a three-dimensional Euclidean space whose dynamics are parameterised by time. The idea of a Minkowski space-time replacing the three dimensions of Euclid and Newton leads to a larger class of symmetries on the underlying space,

instead of geometry being defined by the Euclidean group of translations $\mathbb{R}^3$, reflections and rotations contained in the group $O(3)$, it is replaced by the Poincaré group of four-dimensional translations $\mathbb{R}^{(1,3)}$, and four-dimensional reflections and rotations of the group $O(1,3)$. Three-dimensional vector quantities such as position and momentum are replaced by the position four-vector and the momentum four-vector. These simple ideas lead to the celebrated equation $E = mc^2$, relating energy E to mass m by the large scale factor of the speed of light squared c^2, which, crudely, means that a small amount of mass is equivalent to a large amount of energy. This idea ushered in the nuclear age, in which we now live.

1.2 Maths and Physics

We are studying electromagnetism and special relativity as part of an undergraduate degree in mathematics. You will be forgiven for thinking that the subject sounds much more naturally like a part of a degree in physics. If you have read the brief historic summary above, you will have read about experimental verification of ideas which is of course part of physics. However many of the ideas that guide the discoveries we will study were first made on paper, on the basis of sound mathematical ideas and with only a small input from experimental observation. Such a method of thinking, of using thought experiments rather than actual experiments to guide ideas is central in modern theoretical physics and mathematical physics.

Our aim in this course will be to follow the mathematical ideas that emerge essentially from Coulomb's law of the electrostatic force, the Biot-Savart law for the magnetostatic force and the Lorentz force law and to derive all that follows. Our laboratory will be our minds, together with paper and pen, and we will follow a logical path of mathematical development to see how the more complicated ideas emerge. Many of the consequences of our results will be physically verifiable, however there will be consequences which will shift our mathematical thinking. Chief among these will be the necessity of having to develop a four-dimensional geometry - and the tools needed to work efficiently in four dimensions.

The natural result for a mathematician after having studied this subject is to investigate differential geometry, and see how the geometry of Euclid is generalised. However, without the physical motivation that comes from the unification of electricity and magnetism, the development of higher dimensional geometry in the twentieth century would probably have taken place at a much slower place. The more applied mathematician may well wish to follow this course by studying a course in General Relativity. A second motivation for the mathematician to study this subject is due to the many mathematical tools that are used throughout, this subject is the place to master your knowledge of vector calculus, to begin to think about four-dimensional geometry and to practise applying your knowledge of abstract groups to applied problems. The

subject therefore builds bridges between mathematics and physics, the interplay between these two subjects being the source of many ideas ever since the days of Newton which mark the beginning of the modern era of the scientific method.

1.3 Course Organisation

This year the course will be delivered by live in-person lectures of which there will be three hours per week. Each week in addition to the lectures you should read the corresponding sections in the lecture notes, this weekly reading is an integral part of the course. On the KEATS page for the course you will find a rough outline of the material to be covered each week and in addition a quiz. After you have absorbed the weekly material you should use the quiz to test your knowledge. The quiz is only to help you revise, and you may try each weekly quiz as many times as you wish. No solutions to the qui questions will be posted but if you do not understand the answer to any question please come and discuss in office hours, or use the course discussion board to ask your question. From week two of the teaching semester, there will be an in-person tutorial where problem sets (found at the back of these lecture notes will be discussed) - while particular problems are encouraged, there is no regular marking for the homework you do - instead you should use the tutorials to discuss problems you have encountered. Full solutions to weekly problem sets will be posted on KEATS at the end of the week when the material is covered in the tutorial.

Finally, the lecture notes and the tutorial problems were freshly written. If you find any typos or mistakes please let me know. The lecture notes will reflect exactly what we study in the lectures (and maybe even a little bit more than we study).

1.3.1 Reading List

No extra texts are needed to follow this course. However there are some brilliant books on electromagnetism available, which are difficult to improve upon. Here are those I really enjoy:

1. Introduction to Electrodynamics by David J. Griffiths;
2. The Feynman Lectures on Physics Volume II by Richard Feynman available for reading online at `https://www.feynmanlectures.caltech.edu/II_toc.html`;
3. Lectures on Electromagnetism by Ashok Das;
4. Special Relativity for the Enthusiastic Beginner by David Morin; and
5. Special Relativity by A. P. French.

Only the lecture notes are required to study this course.

2. A Review of Vector Calculus

In which we review the main tools and identities of vector calculus, the language in which electromagnetism will be written. In particular we will review gradient, divergence and curl; line, surface and volume integrals; the fundamental theorem of calculus applied to these kinds of integrals; differential expressions in different coordinate systems; and the Dirac delta function.

This will be covered during week 1 of lectures.

2.1 Vector Derivatives.

The operator $\boldsymbol{\nabla}$ called 'del' or 'nabla' is a vector of partial derivatives. In $\mathbb{R}^3$, $\boldsymbol{\nabla} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix}$ and is used in three different ways, called the gradient, the divergence and the curl. Each one is a form of derivative on either a scalar function or a vector function.[1]

2.1.1 The Gradient.

Definition 2.1.1. *(The gradient) The gradient of a scalar function f is the vector*

$$\boldsymbol{\nabla} f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{pmatrix}.$$

Comment(s). *1. A scalar function is one whose value at any point in $\mathbb{R}^3$ is a number, e.g. air temperature, fluid density, wind speed, altitude etc. See appendix B for a discussion*

[1] We will work in $\mathbb{R}^3$ but each of the following definitions may be generalised to $\mathbb{R}^N$ in a straightforward way, although in the case of the curl the interpretation will be different in dimensions other than three.

of scalars and vectors where they are defined by how they transform under a change of coordinates.

2. *At a point $p \in \mathbb{R}^3$ the direction of $\boldsymbol{\nabla} f$ is determined by the size of its components: it will point in the direction dominated by its largest components, hence it points in the direction that the directional gradient is steepest at p. We can make this argument in another way, let $d\boldsymbol{l} = \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix}$ then*

$$\begin{aligned} \boldsymbol{\nabla} f \cdot d\boldsymbol{l} &= \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz = df \quad \textit{(the total derivative of } f\textit{)} \\ &= |\boldsymbol{\nabla} f||d\boldsymbol{l}| \cos(\theta) \end{aligned}$$

where θ is the angle between $\boldsymbol{\nabla} f$ and $d\boldsymbol{l}$. Let's suppose that $|d\boldsymbol{l}|$ is held fixed and look around the point p to what happens to $\boldsymbol{\nabla} f \cdot d\boldsymbol{l}$ as θ varies. We see that the maximum positive change in f occurs when $\theta = 0$, that is df is maximum when $d\boldsymbol{l}$ points in the direction of $\boldsymbol{\nabla} f$.

Example 2.1. *This is a two-dimensional example, whose benefit is that we can visualise it readily. Consequently $\boldsymbol{\nabla} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix}$. Let $A(x, y) = -x^2 - y^2 + C$, where $C \in \mathbb{R}$ be an altitude function for a hill. The hill has a peak at $(0, 0)$ of height C. The gradient of A at (x, y) is $\boldsymbol{\nabla} A = \begin{pmatrix} -2x \\ -2y \end{pmatrix}$. This is a vector function: for each point of $\mathbb{R}^2$ we have a vector. Apart from at the origin, $\boldsymbol{\nabla} A$ points radially inwards towards the origin (the peak of the hill), i.e. in the direction of steepest ascent at each point. At the origin $\boldsymbol{\nabla} A = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.*

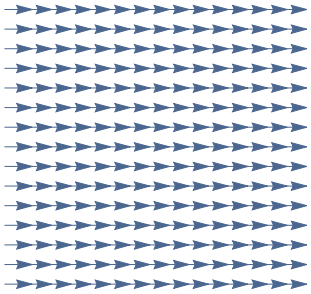
2.1.2 The Divergence.

Definition 2.1.2. *(The Divergence) The divergence of a vector function $\boldsymbol{v}$ is a scalar function given by*

$$\boldsymbol{\nabla} \cdot \boldsymbol{v} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \cdot \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}.$$

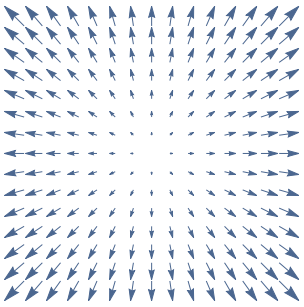
Comment(s). 1. *The divergence is a measure of how quickly the vectors in a vector function change (either in direction of magnitude), in the neighbourhood of a point. Consider the divergence of $\boldsymbol{v}$ in each of the following cases*

- $\boldsymbol{v} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ *which may be sketched by drawing a sample of representative vectors in the* $\boldsymbol{v}$ *at regular intervals:*



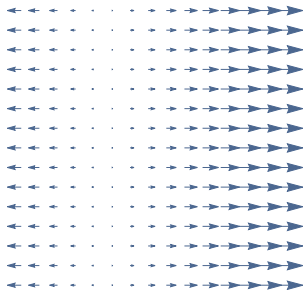
where at each point in $\mathbb{R}^2$ *there is a copy of the vector* $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$. *The vectors all have the same length and do not spread out (or diverge).* $\boldsymbol{\nabla} \cdot \boldsymbol{v} = 0$.

- $\boldsymbol{v} = \begin{pmatrix} x \\ y \end{pmatrix}$:



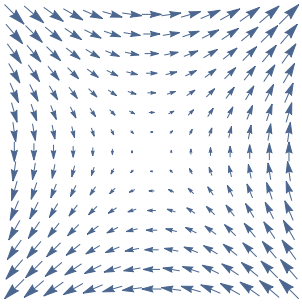
This is a vector field which does diverge everywhere and $\boldsymbol{\nabla} \cdot \boldsymbol{v} = 2$.

- $\boldsymbol{v} = \begin{pmatrix} x+1 \\ 0 \end{pmatrix}$:



This is a vector field in which the vectors remain parallel, but whose lengths differ horizontally in the plane and $\boldsymbol{\nabla} \cdot \boldsymbol{v} = 1$.

- $\boldsymbol{v} = \begin{pmatrix} y \\ x \end{pmatrix}$:

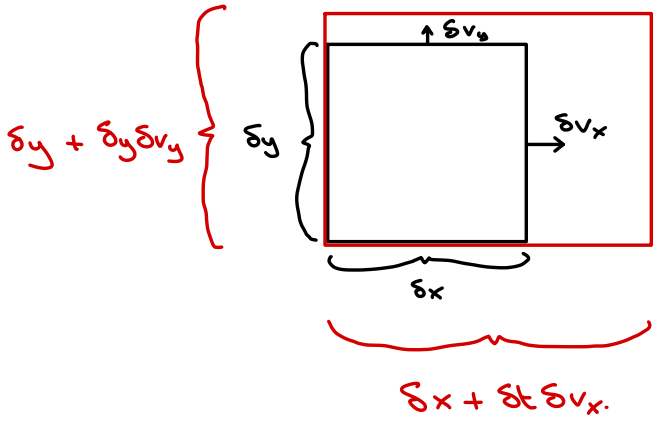


It is not so clear from the sketch to say whether or not this field is diverging: the direction of the arrows appear to get closer together, as we follow the flow of the vectors, but the length of the vectors is increasing. In fact it is divergenceless, the two changes cancelling out: $\boldsymbol{\nabla} \cdot \boldsymbol{v} = 0$.

2. *A nice analogy to help think about the divergence of a vector field is that of fluid velocity, which is a vector field encoding how fast and in which direction a fluid is flowing at each point. In two dimensions, one might be thinking about the velocity vectors at each point on the surface of a pond or river. Now imagine sprinkling confetti onto a small neighbourhood of a point on the surface. If, over time, the confetti spreads out (the area of the patch of confetti grows) then the divergence of the fluid velocity at that point is positive; if it remains unchanged in shape (although it may move) then the divergence is zero; while if the area of the confetti decreases then the divergence is negative. The need to consider a small neighbourhood of a point is important: consider* $\boldsymbol{v} = \begin{pmatrix} x \\ y \end{pmatrix}$ *with* $\boldsymbol{\nabla} \cdot \boldsymbol{v} = 2$, *if this represented a fluid velocity then the point at the origin has* $\boldsymbol{v}(0,0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. *Confetti dropped at the origin would stay put (suggesting zero divergence). A small neighbourhood of the origin, a disc of radius* ϵ *of area* $\pi\epsilon^2$ *grows in time* dt *to a circle of radius*[2] $\epsilon + \epsilon dt$ *so the area of the disc has increased by* $2\epsilon^2\pi dt + \mathcal{O}(dt^2)$, *i.e. it has grown by a factor*

[2]The point $\begin{pmatrix} \epsilon \\ 0 \end{pmatrix}$ on the circumference of the disc has speed ϵ radially outward and so in time dt moves to $\begin{pmatrix} \epsilon + \epsilon dt \\ 0 \end{pmatrix}$. This is similar for all points on the circumference.

of $2dt = \boldsymbol{\nabla} \cdot \boldsymbol{v} dt$. *We can make this comment, this relation between the divergence and the infinitesimal scale factor clear by considering an infinitesimal square centred around a point* p *and allowed to flow as dictated by a vector field* $\boldsymbol{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix}$. *Even though in this picture the two dimensions are flowing under* $\boldsymbol{v}$ *because we consider an infinitesimal square we may choose a frame of reference in which the edges of the square are aligned with the coordinate axes*[3]. *For a sufficiently small square of area* $\delta x \delta y$, *the horizontal change in* v_x, *call it* δv_x, *is approximately constant over the right-hand edge of the square, and similarly the approximate difference in the velocity over the height of the square is* δv_y:



and the sketch above shows the approximate growth of the square in a small time δt. *In* δt *the original area* $\delta x \delta y$ *is changed to*

$$\begin{aligned}(\delta x + \delta t \delta v_x)(\delta y + \delta t \delta v_y) &= \delta x \delta y + (\delta x \delta v_y + \delta y \delta v_x)\delta t + \mathcal{O}(\delta t^2) \\ &= \delta x \delta y \Big(1 + (\frac{\delta v_y}{\delta y} + \frac{\delta v_x}{\delta x})\delta t\Big) + \mathcal{O}(\delta t^2)\end{aligned}$$

Now in the infinitesimal limits where $\delta \to d$, *then the new area becomes*

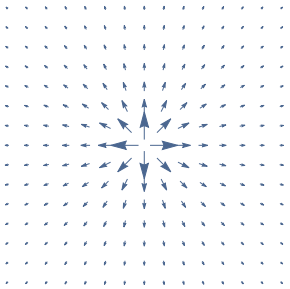
$$dxdy\Big(1 + (\frac{\partial v_y}{\partial y} + \frac{\partial v_x}{\partial x})dt\Big) + \mathcal{O}(dt^2) = dxdy(1 + dt\boldsymbol{\nabla} \cdot \boldsymbol{v}) + \mathcal{O}(dt^2).$$

Therefore $\boldsymbol{\nabla} \cdot \boldsymbol{v}$ *is proportional to the leading term in the change of an infinitesimal area flowing under* $\boldsymbol{v}$.

Example 2.2. *Sketch* $\boldsymbol{v} = \frac{1}{r^2}\boldsymbol{r}$ *where* $\boldsymbol{r} = \begin{pmatrix} x \\ y \end{pmatrix}$ *and compute its divergence. What happens at the origin?*

[3]For a finite square the velocity of points along the edges under $\boldsymbol{v}$ are no longer approximately constant.

The sketch of $\boldsymbol{v}$ *is*



and

$$\begin{aligned}\boldsymbol{\nabla}\cdot\boldsymbol{v} &= \frac{\partial(x/r^2)}{\partial x}+\frac{\partial(y/r^2)}{\partial y}\\ &= \frac{1}{r^2}-2\frac{x}{r^3}\frac{\partial r}{\partial x}+\frac{1}{r^2}-2\frac{y}{r^3}\frac{\partial r}{\partial y}\\ &= 0\end{aligned}$$

where we have used[4] $\frac{\partial r}{\partial x}=\frac{x}{r}$ *and* $\frac{\partial r}{\partial y}=\frac{y}{r}$*. The field is divergence-less, but from the sketch it appears that all points are diverging away from the origin. There is a mismatch between our computation and our interpretation of the sketch (which suggests that the field is diverging at the origin). What has happened is that our computation is not valid at the origin as the field is not well-defined at this point (it involves a division by zero). The message is, as always: beware of division by zero - even at an isolated point. If we were to think about* $\boldsymbol{v}$ *as a fluid velocity then we would see the origin as a point source of fluid flowing into the plane.*

2.1.3 The Curl.

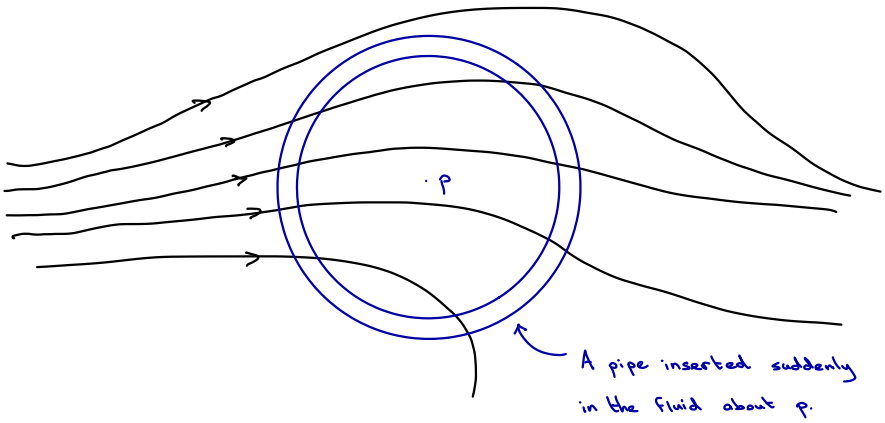
Definition 2.1.3. *(The curl The curl of a vector function* $\boldsymbol{v}$ *is a vector function given by*

$$\boldsymbol{\nabla}\times\boldsymbol{v}=\begin{pmatrix}\frac{\partial}{\partial x}\\ \frac{\partial}{\partial y}\\ \frac{\partial}{\partial z}\end{pmatrix}\times\begin{pmatrix}v_x\\ v_y\\ v_z\end{pmatrix}=\begin{pmatrix}\frac{\partial}{\partial y}(v_z)-\frac{\partial}{\partial z}(v_y)\\ \frac{\partial}{\partial z}(v_x)-\frac{\partial}{\partial x}(v_z)\\ \frac{\partial}{\partial x}(v_y)-\frac{\partial}{\partial y}(v_x)\end{pmatrix}.$$

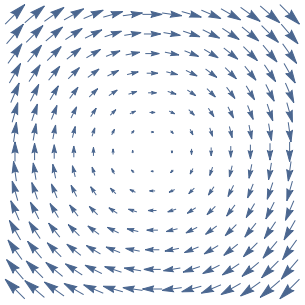
Comment(s). *1. The curl measures how much a vector field circulates about a point. We will make this statement precise in the second comment below, but first let us gain some intuition for what this means. Once again the fluid analogy is useful to obtain a geometric understanding of the curl. Consider a flowing body of water with a smooth velocity vector*

[4] To quickly calculate these partial derivatives act with either $\frac{\partial}{\partial x}$ or $\frac{\partial}{\partial y}$ on $r^2=x^2+y^2$.

field $\boldsymbol{v}$*, where, in the following sketch, we have joined up representative vectors in* $\boldsymbol{v}$ *to indicate the flow-lines or streamlines (where a point flows under* $\boldsymbol{v}$*):*

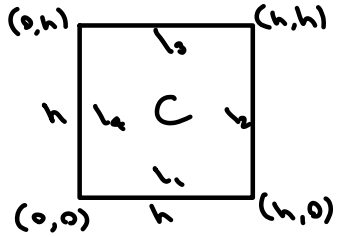


Imagine now that a circular pipe (of infinitesimal radius about p*) could be instantaneously inserted into the fluid around* p*. If the fluid, which is suddenly contained in the pipe, has some circulation around the pipe then the curl of* $\boldsymbol{v}$ *at* p *is non-zero. Consider a simple case where fluid velocity field is* $\boldsymbol{v} = k\begin{pmatrix} y \\ -x \end{pmatrix}$ *where* $k \in \mathbb{R}$ *, then sketching* $\boldsymbol{v}$ *gives*



and $\boldsymbol{\nabla} \times \boldsymbol{v} = \begin{pmatrix} 0 \\ 0 \\ -2k \end{pmatrix}$*. The curl gives us a non-zero whose negative component suggests a clockwise rotation in the plane perpendicular to* $\boldsymbol{\nabla} \times \boldsymbol{v}$*, while a positive component would encode an anti-clockwise rotation under* $\boldsymbol{v}$*. Each interpretation is consistent with using a right-hand rule for* $\boldsymbol{\nabla} \times \boldsymbol{v}$ *to determine the direction of circulation in the plane: to use a right-hand rule, imagine closing the fingers of your right hand with the thumb on the same hand pointing in the direction of* $\boldsymbol{\nabla} \times \boldsymbol{v}$*, then the circulation is in the same direction as your fingers curl when the fist of your right-hand is closed.*

2. *We will make the previous comment precise. However, instead of a circular pipe, we will consider the flow around an infinitesimal square circuit. We will imagine a square of side h with corner at the origin, before taking the limit where $h \to 0$. If $d\boldsymbol{l}$ represents the direction along one of the edges, then the flow of the smooth field $\boldsymbol{v}$ in the direction of $d\boldsymbol{l}$ is $\boldsymbol{v} \cdot d\boldsymbol{l}$. Let us denote the square by C and its boundary by ∂C, then the integral we have argued is related to $\boldsymbol{\nabla} \times \boldsymbol{v}$ is the line integral $\oint_{\partial C} \boldsymbol{v} \cdot d\boldsymbol{l}$. However this integral is proportional to the area of the square h^2, and if we set $h \to 0$ then it vanishes, so instead we are motivated to consider not the circulation but the circulation around the boundary of C per unit area as $h \to 0$, i.e. $\lim_{h\to 0}\left[\frac{1}{h^2}\oint_{\partial C} \boldsymbol{v} \cdot d\boldsymbol{l}\right]$. The contour for this integral is the anti-clockwise path around the square C and we label each leg of the journey as shown on the sketch here:*



Now we may split the line integral around the square into four (ordinary) integrals:

$$\begin{aligned}
\lim_{h\to 0}\Big[\frac{1}{h^2}\oint_{\partial C} \boldsymbol{v}\cdot d\boldsymbol{l}\Big] &= \lim_{h\to 0}\Big[\frac{1}{h^2}\Big(\int_{l_1} \boldsymbol{v}\cdot d\boldsymbol{l} + \int_{l_2} \boldsymbol{v}\cdot d\boldsymbol{l} + \int_{l_3} \boldsymbol{v}\cdot d\boldsymbol{l} + \int_{l_4} \boldsymbol{v}\cdot d\boldsymbol{l}\Big)\Big] \\
&= \lim_{h\to 0}\Big[\frac{1}{h^2}\Big(\int_0^h v_x(x,0)dx + \int_0^h v_y(h,y)dy \\
&\qquad + \int_h^0 v_x(x,h)dx + \int_h^0 v_y(0,y)dy\Big)\Big] \\
&= \lim_{h\to 0}\Big[-\frac{1}{h}\int_0^h \Big(\frac{v_x(x,0+h)-v_x(x,0)}{h}\Big)dx \\
&\qquad + \frac{1}{h}\int_0^h \Big(\frac{v_y(0+h,y)-v_y(0,y)}{h}\Big)dy\Big] \\
&= \lim_{h\to 0}\Big[-\frac{1}{h}\int_0^h \Big(\frac{\partial v_x}{\partial y}\Big)dx + \frac{1}{h}\int_0^h \Big(\frac{\partial v_y}{\partial x}\Big)dy\Big]
\end{aligned}$$

where, in writing the last line above, we have assumed that v_x and v_y admit a Taylor expansion, so that for example $v_x(x,0+h) = v_x(x,0) + h\frac{\partial v_x(x,0)}{\partial y} + \mathcal{O}(h^2)$, hence

$\frac{v_x(x,0+h)-v_x(x,0)}{h} = \frac{\partial v_x}{\partial y} + \mathcal{O}(h)$ *and the error term eventually*[5] *vanishes as*

$$\lim_{h\to 0}\Big[\frac{1}{h}\int_0^h \mathcal{O}(h)dx\Big] = \lim_{h\to 0}\Big[\frac{1}{h}\mathcal{O}(h^2)dx\Big] = 0.$$

Now, for any function $f(t)$ *whose primitive is* $F(t)$ *i.e.* $\frac{dF(t)}{dt} = f(t)$ *then*

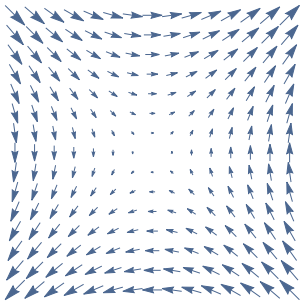
$$\begin{aligned}\lim_{h\to 0}\Big[\frac{1}{h}\int_0^h f(t)dt\Big] &= \lim_{h\to 0}\Big[\frac{1}{h}\Big(F(h)-F(0)\Big)\Big]\\ &= \frac{dF}{dt}\Big|_{t=0}\\ &= f(0).\end{aligned}$$

Returning to our circulation per unit area integral, we see we have

$$\begin{aligned}\lim_{h\to 0}\Big[\frac{1}{h^2}\oint_{\partial C}\boldsymbol{v}\cdot d\boldsymbol{l}\Big] &= \lim_{h\to 0}\Big[-\frac{1}{h}\int_0^h\Big(\frac{\partial v_x}{\partial y}\Big)dx + \frac{1}{h}\int_0^h\Big(\frac{\partial v_y}{\partial x}\Big)dy\Big]\\ &= -\frac{\partial v_x(0,0)}{\partial y} + \frac{\partial v_y(0,0)}{\partial x}.\end{aligned}$$

This is the z-component of $\boldsymbol{\nabla}\times\boldsymbol{v}$ *evaluated at the origin, the point that remained on our small square as we took the infinitesimal limit. In general, the component of* $\boldsymbol{\nabla}\times\boldsymbol{v}$ *normal to a plane is equal to the circulation density per unit area of a vector field around an infinitesimal circuit at* p *in the plane.*

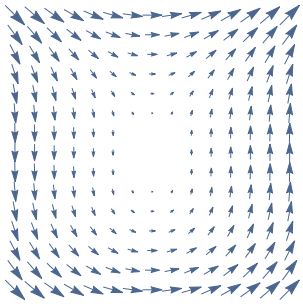
Example 2.3. *Let's consider again the vector field* $\boldsymbol{v} = \begin{pmatrix} y \\ x \end{pmatrix}$*:*



[5]We have suppressed an important detail in the notation here, namely that although the error term is proportional to h it is also a function of x. Our worry is whether the x dependence in $\frac{\partial^2 v_x}{\partial y^2}$ might give a term proportional to $\frac{1}{h}$. However since we have assumed v_x to be a smooth function, then we know its second derivative is continuous. So any small change in its argument x can only induce a small change in its value. As the square is small (at least not infinite) then the changes in values of v_x and its derivatives are finite, and v_x is bounded on the square. But $\frac{1}{h}$ diverges at $h = 0$ on the domain we are integrating, hence v_x and its derivatives cannot be proportional to $\frac{1}{h}$ or any function that grows as fast as $h \to 0$.

Direct computation gives $\boldsymbol{\nabla} \times \boldsymbol{v} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$. *This is a case where the sketch of the vector field may be misleading: from the sketch one might suspect that* $\boldsymbol{\nabla} \times \boldsymbol{v}$ *is in some places non-zero and in the plus or minus z-direction. However the computation shows that this intuition is incorrect. What has gone wrong? Recall that the curl is indicative of the circulation in the plane perpendicular to* $\boldsymbol{\nabla} \times \boldsymbol{v}$*: it must be the case that there is no circulation at any point in the vector field* $\boldsymbol{v}$*. For example, consider the circulation about* $(0, y)$*, the vector field is virtually horizontal (in the x-direction). For a small square about* $(0, y)$ *there is only a contribution to the circulation from the edges of the square parallel to the x-axis. For the edge above* $(0, y)$ $\boldsymbol{v}$ *is larger than along the edge below* $(y, 0)$*, i.e.* $v_x(0, y+\delta) > v_x(0, y-\delta)$ *(for* $y > 0$ *and* $\delta < y$*). So the finite square would have a clockwise circulation around it, but it is the infinitesimal square we are interested in, at which point* $\delta \to 0$ *and we no longer expect any circulation.*

Example 2.4. *Let's consider the vector field* $\boldsymbol{v} = \begin{pmatrix} y^2 \\ x \end{pmatrix}$*:*



Direct computation gives $\boldsymbol{\nabla} \times \boldsymbol{v} = \begin{pmatrix} 0 \\ 0 \\ 1-2y \end{pmatrix}$. *For* $y = \frac{1}{2}$ *the curl is zero, while for* $y < \frac{1}{2}$ $\boldsymbol{\nabla} \times \boldsymbol{v}$ *points in the positive z-direction, so the circulation is in the anti-clockwise sense. For* $y > \frac{1}{2}$ *the circulation is in the clockwise direction.*

The message from these examples is only that intuition about the curl and circulation from sketches of vector fields is difficult, while direct computation of the curl is straightforward.

2.1.4 Product Rules.

As $\boldsymbol{\nabla}$ is constructed from partial derivatives, it inherits the product rules of the derivative.

Gradient Product Rules.

There are two ways to form scalars from products of scalars and vectors, these are fg and $\boldsymbol{u}\cdot\boldsymbol{v}$ where f and g are scalar functions and $\boldsymbol{u}$ and $\boldsymbol{v}$ are vector functions. Then,

$$\boldsymbol{\nabla}(fg) = (\boldsymbol{\nabla}f)g + f(\boldsymbol{\nabla}g) \qquad \text{and} \tag{2.1}$$

$$\boldsymbol{\nabla}(\boldsymbol{u}\cdot\boldsymbol{v}) = \boldsymbol{u}\times(\boldsymbol{\nabla}\times\boldsymbol{v}) + \boldsymbol{v}\times(\boldsymbol{\nabla}\times\boldsymbol{u}) + (\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{v} + (\boldsymbol{v}\cdot\boldsymbol{\nabla})\boldsymbol{u}, \tag{2.2}$$

The way we have written the expression in equation (2.2) is cumbersome but it has the benefit that it is written in terms of standard $\boldsymbol{\nabla}$ and vector products. It is hard to derive this identity, and one helpful method is to work with index notation rather than vector notation - so that the equations become simpler to manipulate. When working with components it is very helpful to invoke a convention called the Einstein summation convention which implies that in any expression where an index label is repeated it is implicit that the index is summed over, i.e. u_iv^i is a short-hand notation for $\sum_{i=1}^{3} u_iv^i$ - we are spared having to write the sigma notation for the sum, and the sum is always over the full range of indices. This notation can be confusing when it is first encountered, and we will use it frequently in this course, so at this point you should read appendix A where all our conventions for index notation will be set up in addition to giving a definition of the Einstein summation convention. Once you have read appendix A, you will be ready to practise using the index notation to prove the results above in the following examples. From hereon in, unless otherwise stated, the Einstein summation convention will be assumed.

Example 2.5. *Use index notation to prove that* $\boldsymbol{\nabla}(fg) = (\boldsymbol{\nabla}f)g + f(\boldsymbol{\nabla}g)$.

As $(\boldsymbol{\nabla})^i = \partial^i$ *then we have*

$$(\boldsymbol{\nabla}fg)^i = \partial^i(fg) = (\partial^i f)g + f(\partial^i g) = (\boldsymbol{\nabla}f)^i g + f(\boldsymbol{\nabla}g)^i.$$

Example 2.6. *Use index notation to prove that*

$$\boldsymbol{\nabla}(\boldsymbol{u}\cdot\boldsymbol{v}) = \boldsymbol{u}\times(\boldsymbol{\nabla}\times\boldsymbol{v}) + \boldsymbol{v}\times(\boldsymbol{\nabla}\times\boldsymbol{u}) + (\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{v} + (\boldsymbol{v}\cdot\boldsymbol{\nabla})\boldsymbol{u}.$$

We will check that the right-hand side equals the left-hand side. In components the left-hand side is

$$(\boldsymbol{\nabla}(\boldsymbol{u}\cdot\boldsymbol{v}))^i = \partial^i(u_jv^j) = (\partial^i u_j)v^j + u_j(\partial^i v^j)$$

which is a much simpler expression, but has the misfortune that it cannot be simply written using standard vector products. We will evaluate each of the terms on the right-hand side of (2.2) *one-by-one. As* $(\boldsymbol{u}\times(\boldsymbol{\nabla}\times\boldsymbol{v}))^n = \epsilon^{nmi}u_m(\epsilon_{ijk}\partial^j v^k)$ *then using* (A.3) *we have*

$$\begin{aligned}(\boldsymbol{u}\times(\boldsymbol{\nabla}\times\boldsymbol{v}))^n &= u_m(\partial^j v^k)\epsilon_{ijk}\epsilon^{inm}\\ &= u_m(\partial^j v^k)(2\delta^{nm}_{jk})\\ &= u_m(\partial^j v^k)(\delta^n_j\delta^m_k-\delta^n_k\delta^m_j)\\ &= u_m(\partial^n v^m)-u_m(\partial^m v^n)\end{aligned}\tag{2.3}$$

The n'th component of the second term on the right-hand side of (2.2) *is the similar to the first term, but with* $\boldsymbol{v}$ *and* $\boldsymbol{u}$ *interchanged. Therefore*

$$(\boldsymbol{v}\times(\boldsymbol{\nabla}\times\boldsymbol{u}))^n = v_m(\partial^n u^m)-v_m(\partial^m u^n).\tag{2.4}$$

The n'th component of the third term in (2.2) *is*

$$((\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{v})^n = u_m\partial^m v^n\tag{2.5}$$

The n'th component of the final term in (2.2) *is*

$$((\boldsymbol{v}\cdot\boldsymbol{\nabla})\boldsymbol{u})^n = v_m\partial^m u^n.\tag{2.6}$$

Adding together equations (2.3)-(2.6) *we obtain*

$$\begin{aligned}u_m(\partial^n v^m)-u_m(\partial^m v^n)+v_m(\partial^n u^m)-v_m(\partial^m u^n)+u_m\partial^m v^n+v_m\partial^m u^n&\\ = u_m(\partial^n v^m)+v_m(\partial^n u^m)&\\ = (\partial^n u_m)v^m+u_m(\partial^n v^m)&\\ = (\boldsymbol{\nabla}(\boldsymbol{u}\cdot\boldsymbol{v}))^n&\end{aligned}$$

as required.

Divergence Product Rules.

There are two ways to form a vector from products of scalars and vectors: $f\boldsymbol{v}$ and $\boldsymbol{u}\times\boldsymbol{v}$ where f is a scalar function and $\boldsymbol{u}$ and $\boldsymbol{v}$ are vector functions. Hence there are two divergence product rules:

$$\boldsymbol{\nabla}\cdot(f\boldsymbol{v}) = f\boldsymbol{\nabla}\cdot\boldsymbol{v}+\boldsymbol{v}\cdot\boldsymbol{\nabla}f\tag{2.7}$$

$$\boldsymbol{\nabla}\cdot(\boldsymbol{u}\times\boldsymbol{v}) = \boldsymbol{v}\cdot(\boldsymbol{\nabla}\times\boldsymbol{u})-\boldsymbol{u}\cdot(\boldsymbol{\nabla}\times\boldsymbol{v})\tag{2.8}$$

Curl Product Rules.

There are also two curl product rules:

$$\boldsymbol{\nabla} \times (f\boldsymbol{v}) = \boldsymbol{\nabla} f \times \boldsymbol{v} + f(\boldsymbol{\nabla} \times \boldsymbol{v}) \tag{2.9}$$

$$\boldsymbol{\nabla} \times (\boldsymbol{u} \times \boldsymbol{v}) = (\boldsymbol{v} \cdot \boldsymbol{\nabla})\boldsymbol{u} - (\boldsymbol{u} \cdot \boldsymbol{\nabla})\boldsymbol{v} + \boldsymbol{u}(\boldsymbol{\nabla} \cdot \boldsymbol{v}) - \boldsymbol{v}(\boldsymbol{\nabla} \cdot \boldsymbol{u}) \tag{2.10}$$

Proofs of equations (2.7)-(2.10) will be done as tutorial exercises.

2.1.5 Second Derivatives.

There are three first derivatives that we may construct using $\boldsymbol{\nabla}$ with scalar (f) and vector ($\boldsymbol{v}$) functions: $\boldsymbol{\nabla} f$ (a vector), $\boldsymbol{\nabla} \cdot \boldsymbol{v}$ (a scalar) and $\boldsymbol{\nabla} \times \boldsymbol{v}$ (a (pseudo)vector). Consequently there are five second derivatives which we may consider:

1. $\boldsymbol{\nabla} \cdot \boldsymbol{\nabla} f$,
2. $\boldsymbol{\nabla} \times \boldsymbol{\nabla} f$,
3. $\boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{v})$,
4. $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{v})$, and
5. $\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{v})$.

These do not all give something new.

1. $\boldsymbol{\nabla} \cdot \boldsymbol{\nabla} f = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial}{\partial x}(f) \\ \frac{\partial}{\partial y}(f) \\ \frac{\partial}{\partial z}(f) \end{pmatrix} = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \partial_i \partial^i f = \boldsymbol{\nabla}^2 f$

$\boldsymbol{\nabla}^2 f$ is called the Laplacian of f. The Laplacian of a scalar function is another scalar function.

2. $\boldsymbol{\nabla} \times \boldsymbol{\nabla} f = 0$ as $\epsilon_{ijk}(\partial^j \partial^k f) = \epsilon_{ijk}(\partial^k \partial^j f) = \epsilon_{ikj}(\partial^j \partial^k f) = -\epsilon_{ijk}(\partial^j \partial^k f)$.
3. $\boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{v}) = \partial^i(\partial_j v^j)$ is not an uninteresting quantity, but it does not show up often in equations in electromagnetism. It is the gradient of the divergence of a vector field, and so is a vector pointing in the direction of fastest increase in the divergences of a vector field.
4. $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{v}) = 0$ as $\partial^i(\epsilon_{ijk}\partial^j v^k) = \epsilon_{ijk}\partial^i \partial^j v^k = -\epsilon_{ijk}\partial^i \partial^j v^k$.

5. $\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{v})$ may be developed in components as follows:

$$\begin{aligned}
\epsilon^{ijk}\partial_j(\epsilon_{klm}\partial^l v^m) &= \epsilon^{ijk}\epsilon_{klm}\partial_j\partial^l v^m \\
&= (\delta^i_l\delta^j_m - \delta^i_m\delta^j_l)\partial_j\partial^l v^m \\
&= \delta^i_l\delta^j_m\partial_j\partial^l v^m - \delta^i_m\delta^j_l\partial_j\partial^l v^m \\
&= \partial_j\partial^i v^j - \partial_j\partial^j v^i \\
&= \partial^i\partial_j v^j - \partial_j\partial^j v^i
\end{aligned}$$

where we have made use of $\epsilon^{ijk}\epsilon_{klm} = \epsilon^{ijk}\epsilon_{lmk} = \delta^i_l\delta^j_m - \delta^i_m\delta^j_l$ as derived and defined in appendix A. This does not have a simple expression in the vector notation. However if we were to interpret $\partial_j\partial^j v^i$ as the Laplacian of a vector as $\boldsymbol{\nabla}^2 = \partial_j\partial^j$ - then we could write the above as the i'th component of $\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{v}) - \boldsymbol{\nabla}^2\boldsymbol{v}$. It may be a little awkward to talk about the Laplacian of a vector quantity, however it is unambiguous. This final second derivative quantity is often used to define precisely what we mean when speak of the Laplacian of a vector:

$$\boldsymbol{\nabla}^2\boldsymbol{v} = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{v}) - \boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{v}). \tag{2.11}$$

For a vector function $\boldsymbol{v} = \begin{pmatrix} v^1 \\ v^2 \\ v^3 \end{pmatrix}$, then $\boldsymbol{\nabla}^2\boldsymbol{v} = \begin{pmatrix} \boldsymbol{\nabla}^2 v^1 \\ \boldsymbol{\nabla}^2 v^2 \\ \boldsymbol{\nabla}^2 v^3 \end{pmatrix}$, it is a vector whose components are the Laplacian acting on the components of $\boldsymbol{v}$.

2.2 Vector Integrals.

2.2.1 The Line Integral.

A line integral is an integral of a field, scalar or vector, along a path. The term *path integral* is often used instead of line integral, while the term *contour integral* is usually reserved for line integrals in the complex plane.

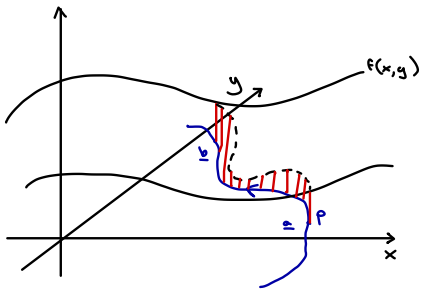
To appreciate the line integral of a vector field, we will first consider the line integral of a scalar field. Consider the simplest case of working in one dimension whose coordinate is x, a real variable, i.e. $x \in \mathbb{R}$. A scalar field in this setting is $f(x)$ and it remains for us to specify a path P (entirely within the x-dimension) over which the integral will be computed. In one dimension any path P from $x = a$ to $x = b$ is imply the interval $[a, b]$, in higher dimensions there will no longer be a unique choice of path. Hence the integral of $f(x)$ over the path $P = [a, b]$

is the standard integral of a function of one (real) variable:

$$\int_P f(x)dl = \int_a^b f(x)dx$$

where $\int_P$ denotes that the integral is a path integral over the path P, and dl is an infinitesimal element along P. Here $dl = dx$ as $P = [a, b]$.

In two dimensions, a scalar function $f(x, y)$ is a function of two real variables denoted here by x and y, so that $\boldsymbol{r} = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2$ and $\int_P f(\boldsymbol{r})dl$ is the integral of $f(\boldsymbol{r})$ along a path P from $\boldsymbol{a}$ to $\boldsymbol{b}$ in $\mathbb{R}^2$. Now $\int_{\boldsymbol{a}}^{\boldsymbol{b}} f(x, y)dl$ is not well-defined (there are an infinite number of paths from $\boldsymbol{a}$ to $\boldsymbol{b}$) unless some variables can be eliminated, and specifying a path P does this for us: along the path x and y are no longer independent variables. An important point to notice here is tht the path P lies in the xy-plane, while $f(x, y)$ may be used to define the height of a surface above the xy-plane: you might think of $f(x, y)$ as an altitude. An attempt to sketch this idea is shown below



where the solid line P is the path from $\boldsymbol{a}$ to $\boldsymbol{b}$ in the xy-plane (and not on the surface defined by $f(x, y)$, the dashed line shows the vertical projection of points on P onto the surface $z = f(x, y)$, and the shaded area bounded above by the dahed line and below by P is the area found by computing the path integral $\int_P f(\boldsymbol{r})dl$ from $\boldsymbol{a}$ to $\boldsymbol{b}$. The path integral in two-dimensions is an integral of a function of one variable, giving an area, and the integral is carried out along the path P which plays the role of the x-axis in the integral of functions of one variable , $f(x)$. We can see specifically the form of the line integral by parameterising P by $t \in [0, 1]$ such that $P(0) = \boldsymbol{a}$ and $P(1) = \boldsymbol{b}$, as then we have

$$\int_P f(\boldsymbol{r})dl = \int_0^1 f(t)\left|\frac{d\boldsymbol{r}}{dt}\right|dt.$$

For higher dimensions the line integral of the scalar functions remains $\int_P f(\boldsymbol{r})dl$ where now $\boldsymbol{r} \in \mathbb{R}^n$ for $n \in \mathbb{Z}^+$, the role of the path P is to eliminate $(n-1)$ variables, so that the line integral is reduced to an integral of a function of one variable along the slice defined by P, which plays the role of the domain of the integral of one variable.

The line integral of a vector field is an integral of the form:

$$\int_P \boldsymbol{v} \cdot d\boldsymbol{l}$$

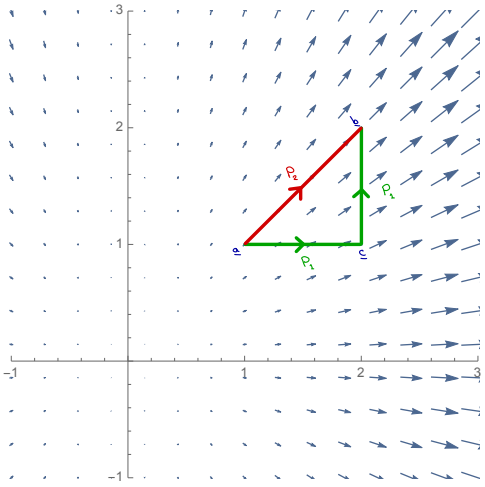
where $\boldsymbol{v}$ is a vector field and $d\boldsymbol{l}$ is an infinitesimal displacement along a path P from $\boldsymbol{a}$ to $\boldsymbol{b}$. If a vector field were everywhere tangential to the path P defining the line integral, then $\boldsymbol{v}(\boldsymbol{r}) \cdot d\boldsymbol{l} = v(\boldsymbol{r})dl$, where $v(\boldsymbol{r}) = |\boldsymbol{v}(\boldsymbol{r})|$ and $dl = |d\boldsymbol{l}|$, and the line integral of the vector field would reduce to a line integral of the scalar field $v(\boldsymbol{r})$. Reversing this argument, one can see that the line integral of a vector field is the line integral of the (scalar) component of $\boldsymbol{v}$ in the direction of the path, $d\boldsymbol{l}$.

Example 2.7. *Calculate the line integral of* $\boldsymbol{v} = \begin{pmatrix} x^2 \\ yx \end{pmatrix}$ *along two different paths from* $\boldsymbol{a} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ *to* $\boldsymbol{b} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ *given by:*

(a) Path 1, P_1*: Travelling from* $\boldsymbol{a}$ *to* $\boldsymbol{b}$ *via* $\boldsymbol{c} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ *along two straight lines.*

(b) Path 2, P_2*: Along the straight line connecting* $\boldsymbol{a}$ *to* $\boldsymbol{b}$

We may sketch the vector field and the paths (although this is not necessary for the computation, it gives us a sense of what we are computing):



The path P_1 *is longer than* P_2*, however the vector field is aligned with* P_2*, so it is difficult to guess which path integral will be the larger. Let's compute:*

(a) Path P_1 *consists of two straight lines, one horizontal from* $\boldsymbol{a}$ *to* $\boldsymbol{c}$*, and the other vertical from* $\boldsymbol{c}$ *to* $\boldsymbol{b}$*. We split the integral along these two paths:*

(i) From $\boldsymbol{a}$ *to* $\boldsymbol{c}$, $d\boldsymbol{l} = \begin{pmatrix} dx \\ 0 \end{pmatrix}$, *from* $x = 1$ *to* $x = 2$, *with* $y = 1$ *all along the path, so that* $\boldsymbol{v}\big|_{y=1} \cdot d\boldsymbol{l} = v^1\big|_{y=1} dx = x^2 dx$.

(ii) From $\boldsymbol{c}$ *to* $\boldsymbol{b}$, $d\boldsymbol{l} = \begin{pmatrix} 0 \\ dy \end{pmatrix}$, *from* $y = 1$ *to* $y = 2$, *with* $x = 2$ *all along the path, so that* $\boldsymbol{v}\big|_{x=2} \cdot d\boldsymbol{l} = v^2\big|_{y=1} dy = 2y dy$.

Altogether we have

$$\begin{aligned} \int_{P_1} \boldsymbol{v} \cdot d\boldsymbol{l} &= \int_1^2 x^2 dx + \int_1^2 2y dy \\ &= \left[\frac{x^3}{3}\right]_1^2 + \left[y^2\right]_1^2 \\ &= \frac{16}{3}. \end{aligned}$$

(b) Path P_2 *lies on the line* $y = x$ *and we use this relation defining the path to eliminate variables until we have just an integral against one variable left, and it does not matter whether we eliminate* x *or* y. *Along* P_2 *we have* $d\boldsymbol{l} = \begin{pmatrix} dx \\ dy \end{pmatrix}$ *hence*

$$\begin{aligned} \int_{P_2} \boldsymbol{v} \cdot d\boldsymbol{l} &= \int_{P_2} \begin{pmatrix} x^2 \\ xy \end{pmatrix} \cdot \begin{pmatrix} dx \\ dy \end{pmatrix} \\ &= \int_{P_2} \left(x^2 dx + xy dy\right) \\ &= \int_1^2 \left(x^2 dx + x^2 dx\right) \\ &= \int_1^2 2x^2 dx \\ &= \frac{14}{3}. \end{aligned}$$

In the third line above we have used $x = y$ *along the path to eliminate* y.

In general the line integral will depend upon the path which defines the integral, although there are vector fields for which the line integral is path-independent as we will see momentarily.

When a line integral is over a closed path i.e. a loop that begins and ends at the same point a special notation is used for the line integral:

$$\oint \boldsymbol{v} \cdot d\boldsymbol{l}$$

the circle on the integral notation indicates that the line integral is taken over a path which is a loop traversed in an anticlockwise sense. For the example above we could form an anticlockwise closed path (or loop) by first travelling P_1 from $\boldsymbol{a}$ to $\boldsymbol{b}$ and then travelling P_2 from $\boldsymbol{b}$ to $\boldsymbol{a}$ i.e.

$$\oint \boldsymbol{v} \cdot d\boldsymbol{l} = \int_{P_1} \boldsymbol{v} \cdot d\boldsymbol{l} - \int_{P_2} \boldsymbol{v} \cdot d\boldsymbol{l} = \frac{16}{3} - \frac{14}{3} = \frac{2}{3}.$$

The minus sign in front of the integral along P_2 occurs because in the closed loop we traverse the path in an anticlockwise fashion, hence path P_2 is travelled from $\boldsymbol{b}$ to $\boldsymbol{a}$, hence the minus sign (as P_2 is defined as a path in the other direction: from $\boldsymbol{a}$ to $\boldsymbol{b}$).

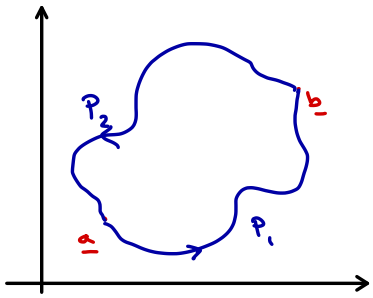
Those vector fields for which $\oint \boldsymbol{v} \cdot d\boldsymbol{l} = 0$ always are rather interesting. Suppose we have such a vector field then we may split the integral around the closed path into two parts

$$\oint \boldsymbol{v} \cdot d\boldsymbol{l} = \int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{v} \cdot d\boldsymbol{l} + \int_{\boldsymbol{b}}^{\boldsymbol{a}} \boldsymbol{v} \cdot d\boldsymbol{l} = 0.$$

So

$$\int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{v} \cdot d\boldsymbol{l} = -\int_{\boldsymbol{b}}^{\boldsymbol{a}} \boldsymbol{v} \cdot d\boldsymbol{l}$$

regardless of the path taken in each line integral between $\boldsymbol{a}$ and $\boldsymbol{b}$:



The only constant information is the beginning and end points, and we note that when we interchange $\boldsymbol{a}$ and $\boldsymbol{b}$ then $\int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{v} \cdot d\boldsymbol{l}$ changes sign. We may hypothesise a simple example of this:

$$\int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{v} \cdot d\boldsymbol{l} = f(\boldsymbol{b}) - f(\boldsymbol{a}).$$

Now if $\boldsymbol{v} = \boldsymbol{\nabla} f$ then

$$\int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{v} \cdot d\boldsymbol{l} = \int_{\boldsymbol{a}}^{\boldsymbol{b}} \begin{pmatrix} \frac{\partial}{\partial x}(f) \\ \frac{\partial}{\partial y}(f) \end{pmatrix} \cdot \begin{pmatrix} dx \\ dy \end{pmatrix} = \int_{\boldsymbol{a}}^{\boldsymbol{b}} df = f(\boldsymbol{b}) - f(\boldsymbol{a}). \tag{2.12}$$

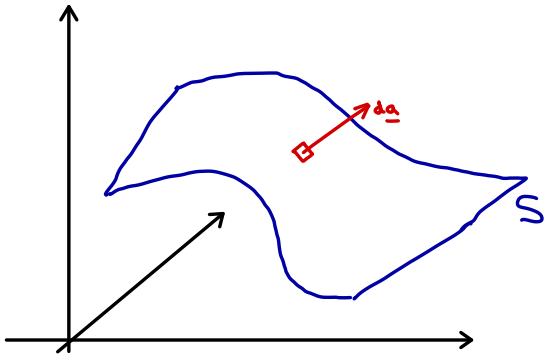
Hence if $\boldsymbol{v} = \boldsymbol{\nabla} f$ then its line integrals are path-independent, and its integrals around a closed path are zero. Such vector fields which are the gradients of a scalar field will be an important class of fields in electromagnetism.

2.2.2 The Surface Integral.

A surface integral is an integral of the form:

$$\int_S \boldsymbol{v} \cdot d\boldsymbol{a}$$

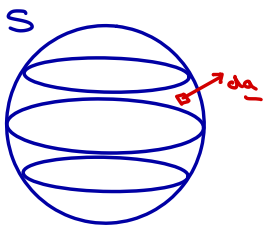
where $\boldsymbol{v}$ is a vector function, $d\boldsymbol{a}$ is an infinitesimal vector whose length is equal to the area of an infinitesimal patch on the surface S and whose direction is perpendicular to the patch of area. The integral is over the whole surface S:



As there are two directions normal to a plane, we have a choice to make. For closed surfaces, ones which entirely bound a volume (e.g. a sphere, an ellipsoid or even an irregular shape forming a balloon around a volume) the two choices of direction for $d\boldsymbol{a}$ correspond to either pointing outwards or inwards from the surface to the interior. When S is a closed surface we write the surface integral of a vector field $\boldsymbol{v}$ over a surface S as

$$\oint \boldsymbol{v} \cdot d\boldsymbol{a}$$

which is the same notation that was used for a line integral about a closed path, except that here we have a surface integral over a closed surface S. The mathematical custom then is to choose the direction of $d\boldsymbol{a}$ so that it points outwards, e.g. for a surface integral over a sphere:



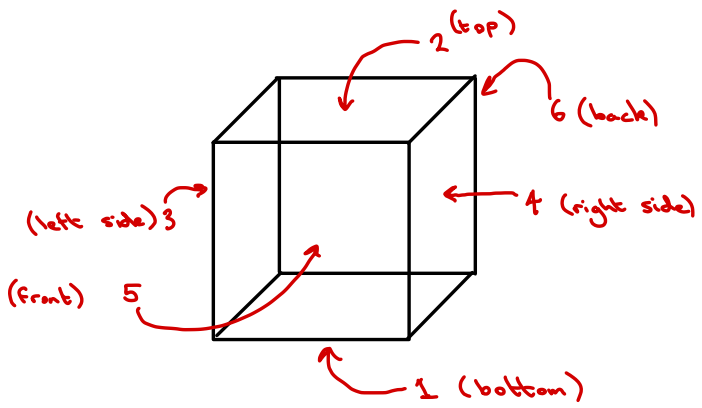
However, for surfaces which are not closed there is no such custom and the choice of direction of $d\boldsymbol{a}$ is arbitrary.

The surface integral sums terms of the form $\boldsymbol{v} \cdot d\boldsymbol{a} = |\boldsymbol{v}||d\boldsymbol{a}| \cos\theta$ which is a product of the size of the component of a vector field perpendicular to a surface element with the area of these elements. This is an unusual quantity to think about geometrically. However when $\boldsymbol{v}$ represents some quantity per unit area, then $\int_S \boldsymbol{v} \cdot d\boldsymbol{a}$ gives the total of that quantity perpendicular to S. A simple example occurs in fluid dynamics when $\boldsymbol{v}$ is taken to be the mass flow of a fluid (or the mass per unit area per unit time), then $\oint \boldsymbol{v} \cdot d\boldsymbol{a}$ is the total mass per unit time crossing the closed surface S: it tells you whether fluid is entering or leaving S. In physics, due to the motion involved, this is called the flux of $\boldsymbol{v}$ through S. Regardless of whether motion is involved, and for whatever vector field $\boldsymbol{v}$ we will also call $\oint \boldsymbol{v} \cdot d\boldsymbol{a}$ the flux of $\boldsymbol{v}$ through S.

The value of a surface integral will normally depend on the shape of S, but in some interesting cases it does not (these is the surface integral equivalent of the path-independent line integral).

Example 2.8. *Calculate the surface integral of* $\boldsymbol{v} = \begin{pmatrix} xy \\ y(z^2-2) \\ y+1 \end{pmatrix}$ *over the surface of the cube of side 2 with vertices at* $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ *and* $\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}$.

The surface of the cube is closed, so we are computing $\oint \boldsymbol{v} \cdot d\boldsymbol{a}$. *On each side of the cube* $d\boldsymbol{a}$ *points in a different direction, so we will compute the flux through each side and add them to obtain the total flux. We will label the sides of the cube as follows:*



Then the surface integrals on each side are

(1) For side 1, $z = 0$ *and* $d\boldsymbol{a} = \begin{pmatrix} 0 \\ 0 \\ -dxdy \end{pmatrix}$ *so* $\boldsymbol{v}|_{z=0} \cdot d\boldsymbol{a} = -v^3|_{z=0} dxdy = -(y+1)dxdy$ *and*

$$\int_{S_1} \boldsymbol{v} \cdot d\boldsymbol{a} = \int_0^2 \int_0^2 -(y+1)dxdy = -8.$$

(2) For side 2, $z = 2$ *and* $d\boldsymbol{a} = \begin{pmatrix} 0 \\ 0 \\ dxdy \end{pmatrix}$ *so* $\boldsymbol{v}|_{z=0} \cdot d\boldsymbol{a} = v^3|_{z=2} dxdy = (y+1)dxdy$ *and*

$$\int_{S_2} \boldsymbol{v} \cdot d\boldsymbol{a} = \int_0^2 \int_0^2 (y+1)dxdy = 8.$$

(3) For side 3, $x = 0$ *and* $d\boldsymbol{a} = \begin{pmatrix} -dydz \\ 0 \\ 0 \end{pmatrix}$ *so* $\boldsymbol{v}|_{x=0} \cdot d\boldsymbol{a} = -v^1|_{x=0} dydz = 0$ *and*

$$\int_{S_3} \boldsymbol{v} \cdot d\boldsymbol{a} = 0.$$

(4) For side 4, $x = 2$ *and* $d\boldsymbol{a} = \begin{pmatrix} dydz \\ 0 \\ 0 \end{pmatrix}$ *so* $\boldsymbol{v}|_{x=2} \cdot d\boldsymbol{a} = v^1|_{x=2} dydz = 2ydydz$ *and*

$$\int_{S_4} \boldsymbol{v} \cdot d\boldsymbol{a} = \int_0^2 \int_0^2 2ydydz = 8.$$

(5) For side 5, $y = 0$ *and* $d\boldsymbol{a} = \begin{pmatrix} 0 \\ -dxdz \\ 0 \end{pmatrix}$ *so* $\boldsymbol{v}|_{y=0} \cdot d\boldsymbol{a} = -v^2|_{y=0} dxdz = 0$ *and*

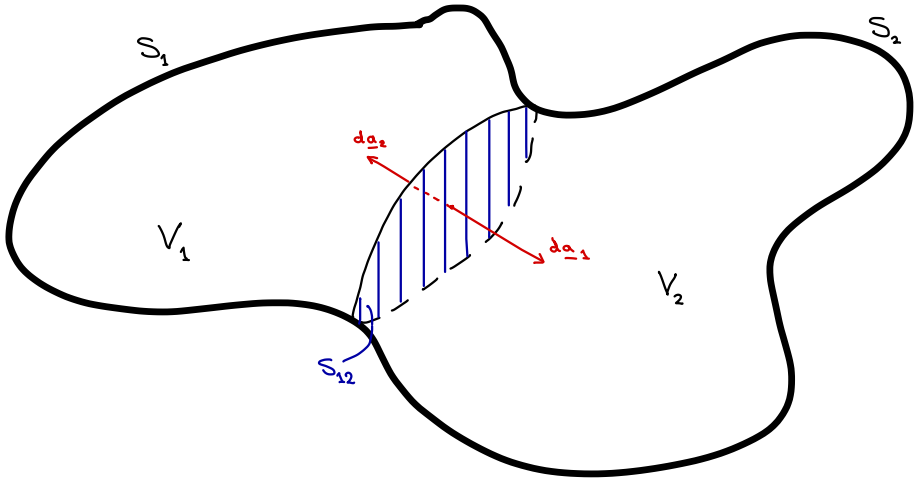
$$\int_{S_5} \boldsymbol{v} \cdot d\boldsymbol{a} = 0.$$

(6) For side 6, $y = 2$ *and* $d\boldsymbol{a} = \begin{pmatrix} 0 \\ dxdz \\ 0 \end{pmatrix}$ *so* $\boldsymbol{v}|_{y=2} \cdot d\boldsymbol{a} = v^2|_{y=2} dxdz = 2(z^2-2)dxdz$ *and*

$$\int_{S_6} \boldsymbol{v} \cdot d\boldsymbol{a} = \int_0^2 \int_0^2 2(z^2-2)dxdz = -\frac{16}{3}.$$

The total flux out of the cube is $\oint \boldsymbol{v} \cdot d\boldsymbol{a} = -8 + 8 + 0 + 8 + 0 - \frac{16}{3} = \frac{8}{3}$. *As the flux is positive, the quantity represented by* $\boldsymbol{v}$ *is leaving the cube on net.*

Consider the computation of the flux through a closed surface S of arbitrary shape (as oppose to the regular cube's surface considered in the example above). Due to the many different directions that the vector $d\boldsymbol{a}$, which is normal to the surface S, will point at different positions on the surface, the general computation of a surface integral looks very challenging. However every surface integral ma be broken down into surface integrals of smaller shapes, and in particular infinitesimal cubes, whose integral will give the surface integral for the shape S. Here we will argue how it is possible to relate the flux through an outer surface, to the flux of many, smaller interior surfaces. A closed surface S bounds a volume V, and we may imagine dividing the volume into two pieces, V_1 and V_2, with two boundaries, sharing a common interior boundary:



We have divided V by taking a slice and creating a new surface S_{12}. The flux of a vector field $\boldsymbol{v}$ through the boundary of V_1 which we will denote as S_{V_1} is

$$\int_{S_{V_1}} \boldsymbol{v} \cdot d\boldsymbol{a} = \int_{S_1} \boldsymbol{v} \cdot d\boldsymbol{a} + \int_{S_{12}} \boldsymbol{v} \cdot d\boldsymbol{a}_1$$

while the flux through the boundary of V_2 which we will denotes by S_{V_2} is

$$\int_{S_{V_2}} \boldsymbol{v} \cdot d\boldsymbol{a} = \int_{S_2} \boldsymbol{v} \cdot d\boldsymbol{a} + \int_{S_{12}} \boldsymbol{v} \cdot d\boldsymbol{a}_2.$$

Now $d\boldsymbol{a}_1 = -d\boldsymbol{a}_2$ so it is clear that if add the fluxes through the two interior surfaces S_{V_1} and S_{V_2} we obtain

$$\int_{S_{V_1}} \boldsymbol{v} \cdot d\boldsymbol{a} + \int_{S_{V_2}} \boldsymbol{v} \cdot d\boldsymbol{a} = \int_{S_1} \boldsymbol{v} \cdot d\boldsymbol{a} + \int_{S_2} \boldsymbol{v} \cdot d\boldsymbol{a} = \int_{S} \boldsymbol{v} \cdot d\boldsymbol{a}$$

where $S = S_1 + S_2$ is the surface of the total volume. In other words the flux of $\boldsymbol{v}$ through the outer surface S can be considered as the sum of the fluxes through all its interior surfaces. It will be natural then to subdivide surfaces of irregular shapes into infinitesimal interior cubes and then integrate the fluxes through all the infinitesimal cubes that make up the irregular volume to find the flux through the larger surface.

2.2.3 The Volume Integral.

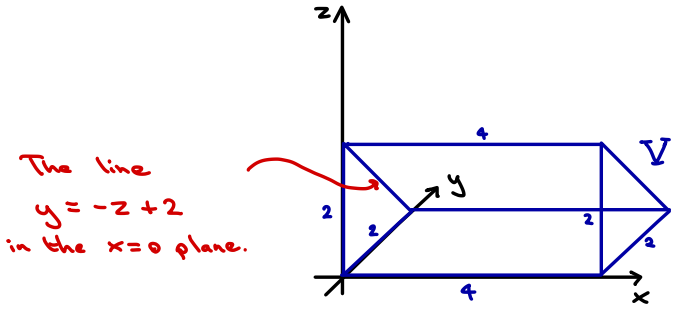
A volume integral is an integral of the form

$$\int_V f dv$$

where f is a scalar function and dv is an infinitesimal volume element in the total volume V. In Cartesian coordinates $dv = dxdydz$. If $f = 1$ the $\int_V dv$ just measures the volume, while if f is a density (a measure of some quantity Q per unit volume) then $\int_V f dv$ gives the total amount of Q contained in V. One can extend the volume integral to a vector field $\boldsymbol{v}$ as

$$\int_V \boldsymbol{v} dv = \begin{pmatrix} \int_v v^1 dxdydz \\ \int_v v^2 dxdydz \\ \int_v v^3 dxdydz \end{pmatrix}.$$

Example 2.9. *Calculate the volume integral of $f = xyz$ over the prism:*



$$\begin{aligned}\int_V f dv &= \int_{x=0}^{x=4} x\Big(\int_{z=0}^{z=2} z\Big(\int_{y=0}^{y=2-z} y dy\Big)dz\Big)dx \\ &= \int_{x=0}^{x=4} x\Big(\int_{z=0}^{z=2} z\Big[\frac{y^2}{2}\Big]_{y=0}^{y=2-z} dz\Big)dx \\ &= \int_{x=0}^{x=4} \frac{x}{2}\Big(\int_{z=0}^{z=2} 4z - 4z^2 + z^3 dz\Big)dx \\ &= \int_{x=0}^{x=4} \frac{2x}{3} dx \\ &= \frac{16}{3}.\end{aligned}$$

2.2.4 The Fundamental Theorems of Vector Calculus

The fundamental theorem of calculus states that

$$\int_a^b f(x)dx = F(b) - F(a)$$

where $\frac{dF}{dx} = f(x)$. The fundamental theorem tells us that the value of an integral, an infinite sum of infinitesimal areas under the curve $y = f(x)$ from $x = a$ to $x = b$, is determined only by values of its primitive $F(x)$ at the end points of the domain: the integral is independent of the interior of the domain.

There are three similar theorems for the integrals of a gradient, a divergence and a curl. Each one relates an integral to the values determined on a boundary of the integral's domain.

The Fundamental Theorem of Calculus for Gradients

We have already shown in equation (2.12) that

$$\int_P (\boldsymbol{\nabla} f) \cdot d\boldsymbol{l} = f(\boldsymbol{b}) - f(\boldsymbol{a}) \tag{2.13}$$

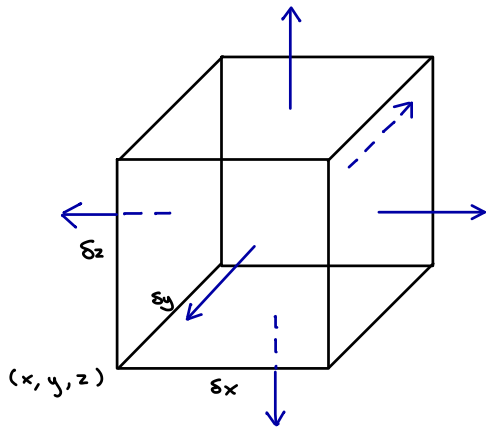
where P is a path from $\boldsymbol{a}$ to $\boldsymbol{b}$ in $\mathbb{R}^3$. This is a generalisation of the fundamental theorem of calculus where the integral is no longer limited to being carried out along the x-axis, instead the line integral is carried out along a one-dimensional path P from $\boldsymbol{a}$ to $\boldsymbol{b}$.

The Fundamental Theorem of Calculus for Divergences (Green's Theorem)

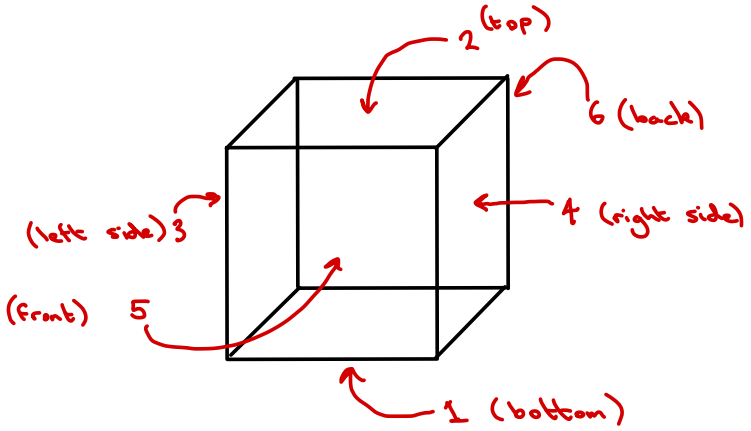
$$\int_V (\boldsymbol{\nabla} \cdot \boldsymbol{u}) dv = \oint_S \boldsymbol{u} \cdot d\boldsymbol{a} \tag{2.14}$$

where $\boldsymbol{u}$ is a vector field, and S is the (closed) boundary of a volume V. This fundamental theorem has many names, it is also known as Gauss's theorem, Green's theorem and also the divergence theorem. It tells us that if we sum all the infinitesimal divergences of a field over a volume, the result is the flux through the outer surface. This result is of no surprise to anyone who has flooded a bathroom: suppose that $\boldsymbol{u}$ is the flow of an incompressible fluid, then the theorem is telling us that the amount of water (the flux of $\boldsymbol{u}$) coming out of the external boundary S is equal to the sum of all the sources and sinks of water throughout the volume V contained by S.

Proof. Consider the small cuboid with one corner located at (x, y, z) and sides of length δx, δy and δz:



We will compute the flux of $\boldsymbol{u}$ through the surface of this cube. To do this we label the faces of the cube as:



First we will compute the flux of $\boldsymbol{u}$ in the x-direction (i.e. through faces 3 and 4). On face 3,

$d\boldsymbol{a} = \begin{pmatrix} -\delta y \delta z \\ 0 \\ 0 \end{pmatrix}$. Since the cube is small, we may approximate the vector field $\boldsymbol{u}$ to be constant over face 3. Then the flux out of face 3 is

$$-u^1(x, y, z)\delta y \delta z,$$

while the flux out of side 4 is

$$u^1(x + \delta x, y, z)\delta y \delta z \approx u^1(x, y, z)\delta y \delta z + \frac{\partial u^1(x, y, z)}{\partial x}\delta x \delta y \delta z.$$

The sum of the fluxes in the x-direction is then approximately

$$\frac{\partial u^1(x, y, z)}{\partial x}\delta x \delta y \delta z.$$

Similarly the total flux in the y-direction is approximated by

$$\frac{\partial u^2(x, y, z)}{\partial y}\delta x \delta y \delta z$$

while the total flux in the z-direction is approximately

$$\frac{\partial u^3(x, y, z)}{\partial z}\delta x \delta y \delta z.$$

Hence,

$$\begin{aligned} \text{total flux} &= \oint_s \boldsymbol{u} \cdot d\boldsymbol{a} \\ &\approx \Big(\frac{\partial v^1(x, y, z)}{\partial x} + \frac{\partial v^2(x, y, z)}{\partial y} + \frac{\partial v^3(x, y, z)}{\partial z}\Big)\delta x \delta y \delta z \\ &= (\boldsymbol{\nabla} \cdot \boldsymbol{u})\delta x \delta y \delta z \end{aligned}$$

In the infinitesimal limit $\delta \to d$, and $\delta x \delta y \delta z \to dv$, the infinitesimal volume, and the approximation becomes an equality:

$$\oint_s \boldsymbol{u} \cdot d\boldsymbol{a} = (\boldsymbol{\nabla} \cdot \boldsymbol{u})dv$$

for the infinitesimal cuboid bounded by s. As we saw earlier, in section 2.2.2, the total flux through a surface is equal to the sum of the fluxes through any complete[6] set of sub-surfaces in its interior. So we may integrate up the infinitesimal flux through the surface of the infinitesimal cuboid to find the flux through a finite surface S:

$$\oint_S \boldsymbol{u} \cdot d\boldsymbol{a} = \int_V (\boldsymbol{\nabla} \cdot \boldsymbol{u})dv.$$

□

[6] A set of surfaces bounding a set of sub-volumes, which together constitute the whole of the volume whose boundary is the original surface.

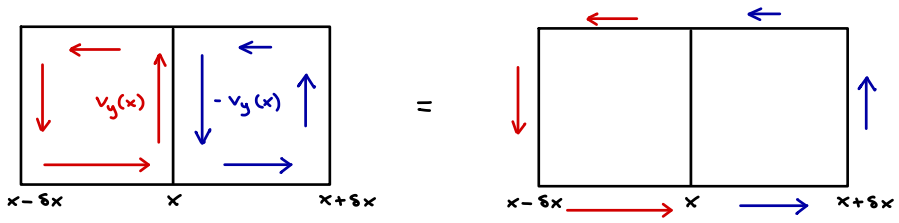
Note that the form of this expression is similar to the form of the fundamental theorem of calculus: the integral over the volume V is determined by data at its boundary S, but a surface has many points and this data is collected into an integral at the surface.

The Fundamental Theorem of Calculus for Curl (Stokes' Theorem)

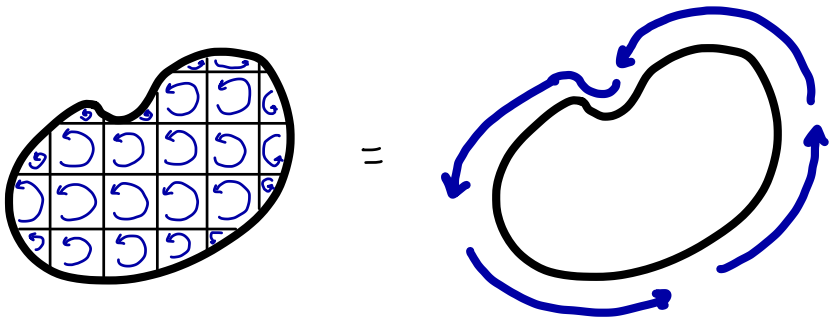
The fundamental theorem of calculus for the surface integral of the curl of a vector function $\boldsymbol{v}$ over an open surface S is

$$\int_S (\boldsymbol{\nabla} \times \boldsymbol{v}) \cdot d\boldsymbol{a} = \oint_P \boldsymbol{v} \cdot d\boldsymbol{l} \tag{2.15}$$

where P is a closed path around the boundary of the surface S. This fundamental theorem is known as Stokes' theorem[7]. Given the interpretation of the $\boldsymbol{\nabla} \times \boldsymbol{v}$ as the infinitesimal circulation density of $\boldsymbol{v}$ and because we defined circulation as the component of a vector field $\boldsymbol{v}$ tangential to a closed path, or loop - perhaps this result is not surprising. Let us connect the dots. We may break a surface up into infinitesimal squares and compute the circulation of a vector field $\boldsymbol{v}$ around each of these smaller squares. As the circulation is measured in an anticlockwise direction, the circulation along all shared boundaries in the lattice of squares will cancel out:



The cancelling out of the circulation at all interior boundaries leaves only the unbalanced circulation at the exterior boundary S:



Just as with the flux, the circulation (or curl) sums over the infinitesimal elements in a surface S to give the total circulation at the boundary of the surface, the path P. This is what Stokes'

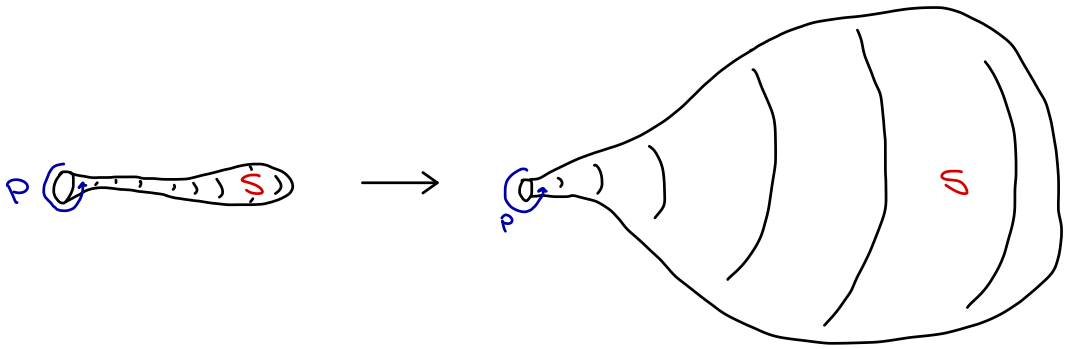
[7]Or sometimes the Kelvin-Stokes' theorem, and sometimes the curl theorem.

theorem is telling us. Intuitively, if a body of fluid is caused to circulate[8] on its surface, then the total circulation would (eventually) be reproduced by the circulation of the fluid at the boundary.

There is one nice and immediate consequence of Stokes' theorem, the value of

$$\int_S (\boldsymbol{\nabla} \times \boldsymbol{v}) \cdot d\boldsymbol{a}$$

depends only on it boundary P, for any vector field $\boldsymbol{v}$. Imagine a surface given by a deflated (open) balloon. Its boundary P is the small circle through which it may be blown up:



Upon blowing up the balloon the surface S become very large, but as the boundary P remains unchanged, then, despite the much large surface S, $\int_S (\boldsymbol{\nabla} \times \boldsymbol{v}) \cdot d\boldsymbol{a}$ is unchanged. Furthermore, if we allow the hole to shrink so that $P \to 0$ then $\oint_P \boldsymbol{v} \cdot d\boldsymbol{l} \to 0$ too. Such a closed surface has no boundary, hence for all closed surfaces we have

$$\oint_S (\boldsymbol{\nabla} \times \boldsymbol{v}) \cdot d\boldsymbol{a} = 0.$$

2.2.5 The Dirac Delta Function

Consider the vector function

$$\boldsymbol{v} = \frac{\boldsymbol{r}}{r^3} = \frac{1}{r^2}\,\boldsymbol{e_r}$$

where $\boldsymbol{r}$ is the radial vector in $\mathbb{R}^3$, $r = |\boldsymbol{r}|$ and $\boldsymbol{e_r} = \frac{\boldsymbol{r}}{r}$. The divergence of $\boldsymbol{v}$ is

$$\boldsymbol{\nabla} \cdot \boldsymbol{v} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \cdot \begin{pmatrix} \frac{x}{r^3} \\ \frac{y}{r^3} \\ \frac{z}{r^3} \end{pmatrix} = 0.$$

[8] Imagine using a thousand teaspoons to stir a cup of tea the size of a swimming pool (why not?), all of the stirring would add up to give the total circulation.

However our intuition about the divergence suggests that this field is spreading out at the origin and should have anon-zero divergence there. Green's theorem (the fundamental theorem for divergences) gives us another way to think about $\boldsymbol{\nabla} \cdot \boldsymbol{v}$. Let's consider the integral over the volume V bounded by the sphere of radius R, the surface S below, centred at the origin:

$$\begin{aligned}\int_V (\boldsymbol{\nabla} \cdot \boldsymbol{v}) dv &= \oint_S \boldsymbol{v} \cdot d\boldsymbol{a} \\ &= \int_S \left(\frac{\boldsymbol{r}}{R^3}\right) \cdot \left((R^2 \cos\phi \, d\theta \, d\phi)\frac{\boldsymbol{r}}{R}\right) \\ &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\phi \cos\phi \int_0^{2\pi} d\theta \\ &= 4\pi\end{aligned}$$

where we have used that $\boldsymbol{r} \cdot \boldsymbol{r} = R^2$ on the surface of the sphere. So the flux of $\boldsymbol{v}$ through the surface of a sphere centred at the origin is 4π regardless of R - we would find the same result if R were extremely small, indicating that the source of the flux is at the origin. On the other hand if $\boldsymbol{\nabla} \cdot \boldsymbol{v} = 0$ throughout the volume bounded by the sphere then Green's theorem is being violated. The problem lies at the origin $\boldsymbol{r} = \boldsymbol{0}$ where both $\boldsymbol{v}$ and $\boldsymbol{\nabla} \cdot \boldsymbol{v}$ are no longer defined, and yet the origin is included within the volume of the integration. We will seek to amend $\boldsymbol{\nabla} \cdot \boldsymbol{v}$ so that it is a generalised function whose value is zero everywhere apart from at the origin, and whose integral is 4π. The generalised function we will need is known as the Dirac delta function.

The one-dimensional Dirac delta function.

In one dimension the Dirac delta function $\delta : \mathbb{R} \to \{0, \infty\}$ is given by

$$\delta(x) = \begin{cases} 0 & \text{if } x \neq 0 \\ \infty & \text{if } x = 0 \end{cases}$$

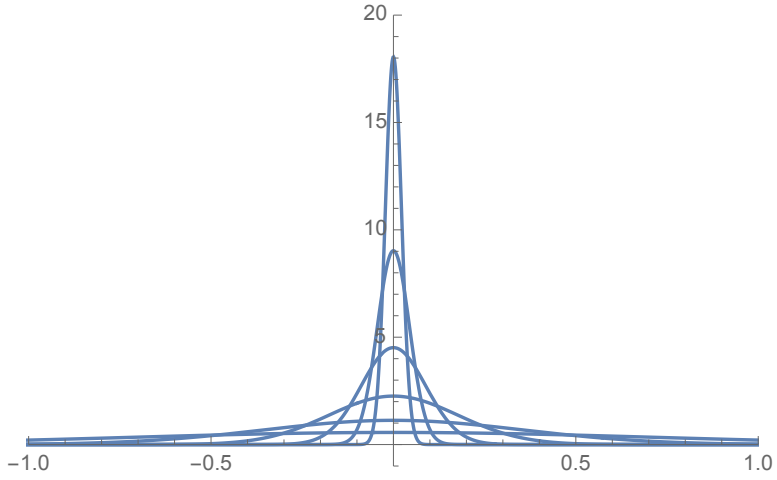
such that

$$\int_{-\infty}^{\infty} \delta(x) dx = 1.$$

Comment(s).

- *The Dirac delta function $\delta(x)$ is an infinitely thin spike located at $x = 0$ but with finite area equal to one.*
- *The Dirac delta function, despite its name, is an example of a generalised function or distribution. It can be understood as the limit of a sequence of functions, for example,*

let $\delta_n(x) = \frac{1}{|n|\sqrt{\pi}} e^{-(\frac{x}{n})^2}$, *then* $\delta(x) = \lim_{n\to 0}(\delta_n(x))$. *In the sketch below you can see the functions plotted around* $x = 0$ *for* $n = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$, *and* $\frac{1}{32}$.



- *The Dirac delta function is a continuous version of the Kronecker delta function* δ^i_j *for which* $\delta^i_j v^j = v^i$, *the corresponding formula for the Dirac delta function involves an integral:*

$$\int_{-\infty}^{\infty} \delta(x) f(x) dx = \int_{-\infty}^{\infty} \delta(x) f(0) dx = f(0) \int_{-\infty}^{\infty} \delta(x) dx = f(0)$$

where we have used the fact that the Dirac delta function is zero for all $x \neq 0$ *in order to replace* $f(x)$ *with* $f(0)$ *under the integral. Equivalently,*

$$\int_{-\infty}^{\infty} \delta(x-a) f(x) dx = \int_{-\infty}^{\infty} \delta(x-a) f(a) dx = f(a) \int_{-\infty}^{\infty} \delta(y) dy = f(a).$$

The three-dimensional Dirac delta function.

The three-dimensional Dirac delta function is denoted δ^3 and maps $\mathbb{R}^3 \to \{0, \infty\}$, it is given by

$$\delta^3(\boldsymbol{r}) = \delta(x)\delta(y)\delta(z)$$

where $\boldsymbol{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$. Now,

$$\int_{\mathbb{R}^3} \delta^3(\boldsymbol{r}) dv = \int_{-\infty}^{\infty} \delta(z) dz \int_{-\infty}^{\infty} \delta(y) dy \int_{-\infty}^{\infty} \delta(x) dx = 1.$$

We also have,

$$\int_{\mathbb{R}^3} \delta^3(\boldsymbol{r} - \boldsymbol{a}) f(\boldsymbol{r}) dv = f(\boldsymbol{a}).$$

Returning to $\boldsymbol{v} = \frac{\boldsymbol{r}}{r^3}$ we may resolve the conundrum of how to define $\boldsymbol{\nabla} \cdot \boldsymbol{v}$ so that it has a contribution to the flux through a spherical shell about the origin of 4π. If we set

$$\boldsymbol{\nabla} \cdot \boldsymbol{v} = 4\pi\delta^3(\boldsymbol{r})$$

then we have a function which is zero everywhere, but infinite at the origin. But under the integral over the volume bounded by the sphere S^2, denoted by V below, we have a contribution at $\boldsymbol{r} = \boldsymbol{0}$:

$$\int_V (\boldsymbol{\nabla} \cdot \boldsymbol{v}) dv = \int_V (4\pi\delta^3(\boldsymbol{r})) dv = 4\pi$$

matching the flux integral it is related to by Green's theorem: $\oint_S \boldsymbol{v} \cdot d\boldsymbol{a} = 4\pi$.

We will introduce some notation that we will use frequently throughout the course, let $\boldsymbol{s} = \boldsymbol{r} - \boldsymbol{r'}$ denote a separation vector (from $\boldsymbol{r'}$ to $\boldsymbol{r}$) where $\boldsymbol{r}$ is a variable, but $\boldsymbol{r'}$ is a constant vector. Then,

$$\boldsymbol{\nabla} \cdot \left(\frac{\boldsymbol{s}}{s^3}\right) = 4\pi\delta^3(\boldsymbol{s})$$

where $s^2 = (x-x')^2+(y-y')^2+(z-z')^2$, and further we may also deduce that since $\boldsymbol{\nabla}(\frac{1}{s}) = -\frac{\boldsymbol{s}}{s^3}$ then

$$\boldsymbol{\nabla}^2\left(\frac{1}{s}\right) = -4\pi\delta^3(\boldsymbol{s}). \tag{2.16}$$

3. Electrostatics

Electrostatics is a simplified version of electromagnetism and is the study of systems of electric charges which are at rest. Electric charge is carried by many fundamental particles, and all elementary particles have charges which are integer multiples of $\frac{1}{3}$e where e is the magnitude of the charge of the electron and, is roughly $1.602176638 \times 10^{-19}$ Coulombs. The charge on the electron is negative, it is $-e$, while the charge on the proton is positive $+e$. We will work with mathematical charges whose sizes we will denote q or Q and our first concern will simply be whether they are positive or negative, rather than how small or large they are.

Before entering this chapter proper, it's a good idea to familiarise yourself with Appendix C, where our conventions for cylindrical and spherical coordinates are detailed. In particular note that hats denote unit vectors throughout and basis vectors may be denoted by $\boldsymbol{e_r}$ or just by $\hat{\boldsymbol{r}}$, where it is convenient, and similarly for other coordinates.

3.1 Coulomb's Law.

Experiment has shown that the force on a test charge Q at $\boldsymbol{r}$ due to a single point charge q at $\boldsymbol{r'}$ is given by Coulomb's Law:

$$\boldsymbol{F} = \frac{1}{4\pi\epsilon_0}\frac{qQ}{s^2}\hat{\boldsymbol{s}} \tag{3.1}$$

where $\boldsymbol{s} = \boldsymbol{r} - \boldsymbol{r'}$ is the displacement vector, $\hat{\boldsymbol{s}} = \frac{\boldsymbol{s}}{s}$ is the unit displacement vector, and ϵ_0 is a constant called the permittivity of free space. The force $\boldsymbol{F}$ is measured in Newtons (N), the displacement in metres (m)and the charges q and Q in Coulombs (C), so that

$$\epsilon_0 \approx 8.85418782 \times 10^{-12}\frac{C^2}{Nm^2}.$$

The permittivity of free space allows us to use the standard (S.I.)[1] units to measure force, charge and distance in this equation. There is another commonly used set of units for studying

[1]From the French 'Système international (d'unités)', or the International System of Units.

electromagnetism called Gaussian units which are part of the C.G.S.[2] set of units. The unit of measurement for charge in Gaussian units is the statcoulomb, abbreviated statC, where 1 statC $=1\ g^{\frac{1}{2}}cm^{\frac{3}{2}}s^{-1}$, as distance is measured in cm then the unit of force is measured in another new unit the dyne. The advantage of Gaussian units is that the set of constants of proportionality frop out of Coulomb's law which in Gaussian (or C.G.S.) units is

$$\boldsymbol{F} = \frac{qQ}{s^2}\hat{\boldsymbol{s}}.$$

However throughout this course we will adopt the S.I. units and make do with the factors of $\frac{1}{4\pi\epsilon_0}$.

The Coulomb force is proportional to the charges, inversely proportional to the distance squared between the charges and is in the direction of the straight line between the two charges. Notice that if both Q and q are positive or both are negative then the force on Q points in the direction of $\boldsymbol{s} = \boldsymbol{r} - \boldsymbol{r'}$, which points from q to the test charge Q, i.e. if the two charges have the same sign then the Coulomb force is repulsive. While if the two charges have opposite signs, then the Coulomb force is attractive. The Coulomb force acts on the test charge Q and is written in terms of the separation vector $\boldsymbol{s}$ from $\boldsymbol{r'}$ to the position of the test charge $\boldsymbol{r}$. The fact that the equation is a vector equation (i.e. both sides of the equations are vectors) means that both sides transform in the same way under a coordinate transformation, and so if we wish we could always translate the coordinates so that the source charge q sits at the origin $\boldsymbol{0}$ instead of $\boldsymbol{r'}$, in this case the equation keeps the same form but $\boldsymbol{s} \to \boldsymbol{r}$ so that the Coulomb force equation becomes $\boldsymbol{F} = \frac{1}{4\pi\epsilon_0}\frac{qQ}{r^2}\hat{\boldsymbol{r}}$ which might be a more familiar formulation.

The ideas above bears further investigation. Two charges q and Q exert a force on each other, and the equation for the Coulomb force tells us only what the instantaneous force on the test particle is. With this equation we can imagine computing the force on Q due to the point charge q at all points in space. We would say that the force $\boldsymbol{F}$ is a vector field that defines a vector for all points in space. We might start to wonder about the physical existence of this field due to a point charge q, in particular we may ask whether it is a property of the point particle, rather than a force which exists only when a test particle is present. The reality of the force field is something we will think about and probe throughout the course.

How does the point charge q bring a force to act on the test charge Q? It may bother us that there is no obvious medium for the transfer of momentum, it may also bother us if we start to think about when does the force act? In the Coulomb law the force is acting instantaneously over a distance s, a so-called action at a distance. The phrase action at a distance is something that sounds worrying at the human level of the natural world. We are accustomed to forces acting by objects colliding with each other and momentum being transferred. This is something

[2]Centimetre-gram-second units.

of a red herring, although we can't see it with our eyes, when two macroscopic objects collide, it is only the repulsion of the negatively charged electrons in the outer shells of the surface atoms that is driving the dynamics. The electrons never touch and bounce off each other, rather it is the action at a distance of the electromagnetic force that causes the momentum to be transferred. Our intuition that action at a distance is unnatural must give way to the notion that all forces at the human level of the world are happening via some action at a distance. It is the most natural kind of force, it is just that the distances are so small that our eyes do not see them. Nevertheless we may still be bothered that somehow information such as the amount of charge q has to be communicated from $\boldsymbol{r'}$ to $\boldsymbol{r}$ in order to inform the test charge Q how to move. Since the distance between the two particles may be arbitrarily large - as large say as the width of the known universe, is it really viable that the force would act instantaneously? The answer will be that the force travels at the speed of light, and it is not instantaneous after all - we will need Maxwell's equations to see this. However, as a result, the force a test charge Q feels will depend on the position of the point charge q in the past, just far enough in the past for a signal at the speed of light to traverse the distance between q and Q.

Another thought we may have about the Coulomb force is how would we ever be able to use it? As it tells us about an unlikely scenario: it tells us about the instantaneous force on a charge Q which is, initially, stationary with respect to the point charge q. Notice that the strength of the Coulomb force diminishes as $\frac{1}{s^2}$, so that it is present throughout all space: if that is the case how is it possible that we could ever arrange to bring a test charge Q to a distance s of the point charge q without disturbing it and causing it to move. This problem may not bother us deeply, we can think of experimental arrangements that can be set up to test the idea, which would hold q and Q fixed until we are ready to measure $\boldsymbol{F}$. But the impracticality of the arrangement hints that this force law will need to be generalised to take account of charges which are in motion with respect to each other. This is what Maxwell's equations will eventually do for us, among other things.

However this unlikely situation of systems of charged particles being held still so that we may describe the instantaneous force a test charge Q would feel, will greatly inform us about the path to Maxwell's equations. This test setting is called electrostatics - for it involves static arrays of electric charges. How we may use Coulomb's law together with vector calculus to work out the force exerted by an array of electric charges, in many interesting ways.

3.2 The Electric Field

Coulomb's law is a vector equation, so we may add the forces[3] due to several source charges q_1, q_2 $\ldots q_n$ at displacements (from q_i to Q) given by $\boldsymbol{s_1} = \boldsymbol{r} - \boldsymbol{r_1'}$, $\boldsymbol{s_2} = \boldsymbol{r} - \boldsymbol{r_2'}$, $\ldots \boldsymbol{s_n} = \boldsymbol{r} - \boldsymbol{r_n'}$ to find the total force on a test charge Q

$$\boldsymbol{F} = \frac{1}{4\pi\epsilon_0}\Big(\frac{q_1 Q}{s_1^2}\hat{\boldsymbol{s_1}} + \frac{q_2 Q}{s_2^2}\hat{\boldsymbol{s_2}} + \ldots \frac{q_n Q}{s_n^2}\hat{\boldsymbol{s_n}}\Big) = Q\boldsymbol{E}(\boldsymbol{r})$$

where

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\sum_{i=1}^{n}\frac{q_i}{s_i^2}\hat{\boldsymbol{s_i}}$$

is called the electric field of the source charges. It is independent of the test charge, and it is really the physical reality of the electric field which we may wonder about: we only know about it through the force $\boldsymbol{F} = Q\boldsymbol{E}(\boldsymbol{r})$ it induces on a test charge Q.

We may readily generalise our definition of the electric field $\boldsymbol{E}$ due to a collection of point charges to a continuous distribution of charge over a volume. Let $\rho(\boldsymbol{r'})$ denote the charge per unit volume, or charge density for a continuous distribution of charges. Then the electric field is the integral

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r'})}{s^2}\hat{\boldsymbol{s}}dv'. \tag{3.2}$$

In evaluating the electric field for a continuous charge distribution, one must keep in mind that we are integrating to find the total source charge, hence the integral is over the volume containing the charges - which in the continuous limit is now $\boldsymbol{r'}$ rather than $\boldsymbol{r_i'}$. We emphasise that $\boldsymbol{r}$, the position of a test charge Q is constant in the integral. This is why we have denoted the volume element above as dv', rather than dv, as it is an infinitesimal volume about $\boldsymbol{r'}$. For charge distributions restricted to a surface or to a line, one may find the electric field by evaluating the appropriate surface or line integral.

In fact by using the Dirac delta function we may also construct a charge density for a discrete charge q_1 at the position $\boldsymbol{r_1'}$, it is

$$\rho(\boldsymbol{r}) = q_1\delta^3(\boldsymbol{r} - \boldsymbol{r_1'})$$

[3]Experimental tests confirm this.

for which

$$\begin{aligned}
\boldsymbol{E}(\boldsymbol{r}) &= \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r'})}{s^2}\hat{\boldsymbol{s}}dv' \\
&= \frac{1}{4\pi\epsilon_0}\int \frac{q_1\delta^3(\boldsymbol{r'}-\boldsymbol{r'_1})}{s^2}\hat{\boldsymbol{s}}dv' \\
&= \frac{1}{4\pi\epsilon_0}\int \frac{q_1\delta^3(\boldsymbol{r'}-\boldsymbol{r'_1})}{|\boldsymbol{r}-\boldsymbol{r'}|^2}\frac{\boldsymbol{r}-\boldsymbol{r'}}{|\boldsymbol{r}-\boldsymbol{r'}|}dv' \\
&= \frac{1}{4\pi\epsilon_0}\frac{q_1}{|\boldsymbol{r}-\boldsymbol{r'_1}|^3}(\boldsymbol{r}-\boldsymbol{r'_1}) \\
&= \frac{1}{4\pi\epsilon_0}\frac{q_1}{s_1^2}\hat{\boldsymbol{s}_1}
\end{aligned}$$

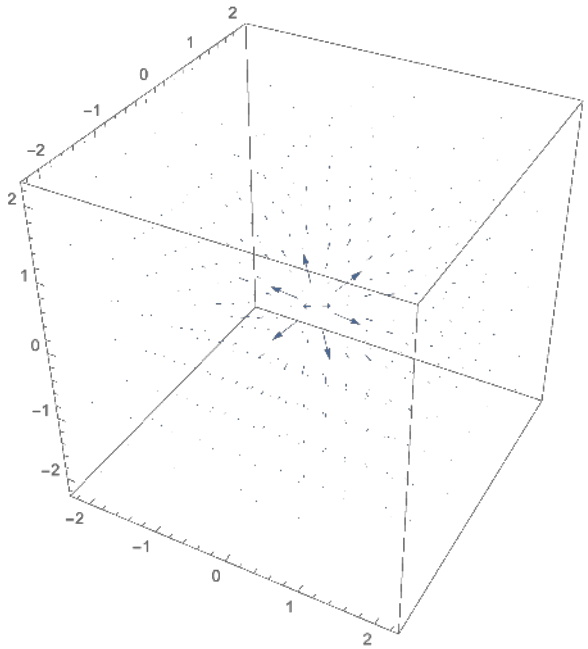
as anticipated.

3.2.1 Electric Field Lines

One way to picture an electric field is to sketch it as a vector field, drawing a set of representative vectors for the field. The electric field for a single point charge q at the origin is

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\boldsymbol{e_r} \tag{3.3}$$

where $\boldsymbol{e_r} = \hat{\boldsymbol{r}}$. This is proportional to the vector field $\boldsymbol{v} = \frac{\boldsymbol{r}}{r^3}$ whose divergence we thought about a little in section 2.2.5. The vectors point radially outward from the origin, in three dimensions the vector field is difficult to comprehend:



Instead we sketch the two-dimensional cross-section of the field:



However we also see that the size of the vectors (varying as $\frac{1}{r^2}$) makes a meaningful plot difficult to render. Instead what we may do is to represent the field in terms of its field lines, these are the lines one would find if we joined up the vectors in our sketch head to tail to form lines. There are some rules to drawing electric field lines which we will outline here

(i) Field lines are continuous curves whose tangents at any point $\boldsymbol{r}$ are parallel to $\boldsymbol{E}(\boldsymbol{r})$, and have the same direction as $\boldsymbol{E}(\boldsymbol{r})$

(ii) When sketching field lines only a finite number of curves are drawn, whose density is representative of the strength of the vector field.

(iii) All field lines must begin on either a positive charge or at infinity and end on a negative charge or at infinity.

(iv) The number of lines emanating from a point source is proportional to the charge of the source and symmetrically distributed around it.

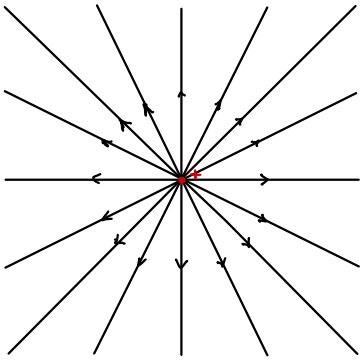
Comment(s).

- *Field lines are normally drawn for a two-dimensional cross-section of a three-dimensional field.*

- *Field lines point away from positive charges and towards negative charges, they have the direction of the force on a positive test charge.*

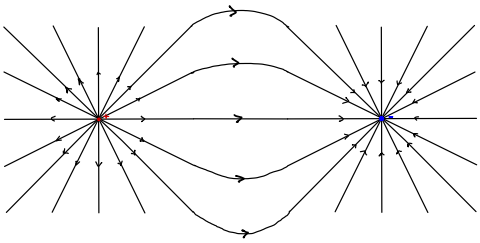
- *Electric field lines do not intersect each other: if they did it would mean that $\boldsymbol{E}(\boldsymbol{r})$ was badly defined (having two vectors at a single point).*

- *The electric field lines for a point charge is*



- *The density of field lines is representative of the strength of* $\boldsymbol{E}(\boldsymbol{r})$*: for the point charge we see that the field lines are dense near the charge and goes down as the lines spread out further from the origin. If we sketch* n *field lines then the density of lines on a circle of radius* r *is* $\frac{n}{2\pi r} \propto \frac{1}{r}$*, and not* $\frac{1}{r^2}$ *as claimed. However we are sketching a cross-section of a three-dimensional field, so the density of field lines should be computed using a sphere of radius* r*, in which case the density for* n *emanating lines is* $\frac{n}{4\pi r^2} \propto \frac{1}{r^2}$ *as claimed.*

- *It can sometimes be very simple to intuit the field lines for a static array of charges, as between symmetry, the rules for drawing the field lines, and some rough knowledge of the electric field, almost everything is determined. For example consider the field for two point charges, one positive and one negative. In the vicinity of each charge we should find field for a point charge. Sketching the rest of the field is then about consistently joining up these lines bearing in mind that field lines cannot cross:*



- *The flux of* $\boldsymbol{E}(\boldsymbol{r})$ *through a surface is a measure of the number of field lines drawn cutting that surface: since the density of field lines (the number per unit area) is proportional* $\boldsymbol{E}$ *then* $\boldsymbol{E} \cdot d\boldsymbol{a}$ *in the computation of the flux is proportional to the number of lines cutting the surface. Of course since we only sketch a representative set of field lines, the actual number for any field lines on a given sketch is meaningless, but is representative of the flux.*

3.3 Gauss's Law

Our rules for sketching field lines are suggestive. If field lines can only begin or end on a source charge or at infinity, then the number of field lines entering or leaving any bounded finite region will depend upon the amount of charge contained in the bounded region. Source charges are sources or sinks for field lines, hence field lines from a positive charge must either pass through some bounding surface or must end on a negative charge within the surface. Let's calculate the flux for the electric field for a point charge located at the origin as given in equation (3.3) through a spherical surface centred at the origin and of radius R:

$$\begin{aligned}\oint_S \boldsymbol{E}(\boldsymbol{r})\cdot d\boldsymbol{a} &= \int_{\phi=-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_{\theta=0}^{2\pi}\Big(\frac{1}{4\pi\epsilon_0}\frac{q}{R^2}\,\boldsymbol{e_r}\Big)\cdot(R^2\cos(\phi)d\theta d\phi\;\boldsymbol{e_r})\\ &= \frac{q}{4\pi\epsilon_0}\int_{\phi=-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_{\theta=0}^{2\pi}\cos(\phi)d\theta d\phi\\ &= \frac{q}{\epsilon_0}.\end{aligned}$$

The radius of the sphere cancels out[4] because it grows as r^2 while the electric field decreases as $\frac{1}{r^2}$. Since the flux is proportional to the number of field lines cutting a surface, then this is reassuring: the same number of field lines around a single source charge will cut a sphere of small radius as a sphere of large radius (no field lines are created or destroyed because there are no other charges present). We made the argument when the closed surface S is a sphere, however we would get the same result for any closed surface S. For any other surface S the coordinates θ and ϕ will still vary over the solid angle of 4π, but the value of r would vary as the integral goes over the surface - however no matter what value r^2 takes in the integral at any point in the surface it will still be cancelled out by the $\frac{1}{r^2}$ in $\boldsymbol{E}(\boldsymbol{r})$. If we now consider the electric field for a point charge located at $\boldsymbol{r'}$, as opposed to the origin, then the flux through any closed surface S containing the charge is still $\frac{q}{\epsilon_0}$.

Furthermore we know that the electric field for multiple point sources, is found by adding the vector fields for each source, hence

$$\oint_S \boldsymbol{E}(\boldsymbol{r})\cdot d\boldsymbol{a} = \oint_S\Big(\sum_{i=1}^{n}\boldsymbol{E_i}(\boldsymbol{r})\cdot d\boldsymbol{a}\Big) = \sum_{i=1}^{n}\Big(\oint_S \boldsymbol{E_i}(\boldsymbol{r})\cdot d\boldsymbol{a}\Big) = \sum_{i=1}^{n}\frac{q_i}{\epsilon_0}$$

If we denote the total charge contained within the surface S by $Q_S = \sum_{i=1}^{n} q_i$ then

$$\oint_S \boldsymbol{E}(\boldsymbol{r})\cdot d\boldsymbol{a} = \frac{Q_S}{\epsilon_0}. \tag{3.4}$$

[4]This would be the same for the Newton's theory of gravity, where there is a corresponding gravitational version of Gauss's law.

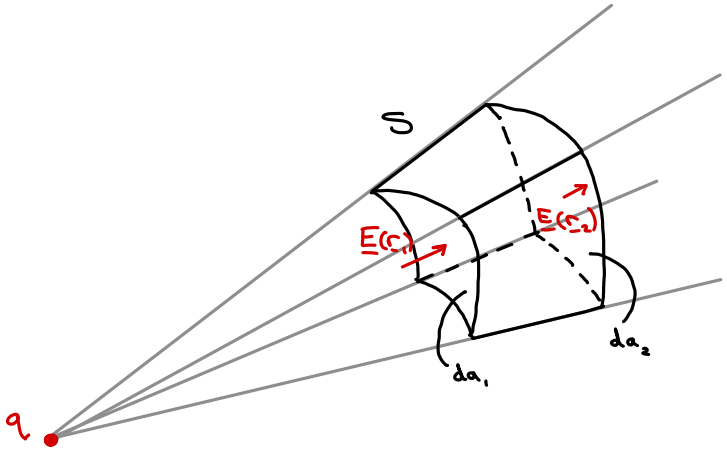
Equation (3.4) is Gauss's Law.

Comment(s).

- *Gauss's Law tells us that the flux through any closed surface S, no matter its shape, is proportional to the sum of the charges enclosed within the surface. It is insensitive to where the source charges are located, as well as to the shape of S. For example if we imagined a continuous distribution of charges spread uniformly over the volume of a ball centred at the origin and within S whose total charge is Q_S, then outside the ball the flux of the electric field is the same as that for a point source of charge Q_S located at the origin*[5].

- *If the surface in the flux integral contains net zero charge then the flux through the surface is zero.*

- *After some thought Gauss's law looks too good to be true: complicated integrals are being vastly simplified. We may start to wonder how can it be that a bounding surface containing no charge, but in an electric field, which will have field lines cutting it (in a way that will depend on the shape of S) have zero flux? Consider the point source and now let S be a small curving cuboid surface running radially (in spherical coordinates) from $r = r_1$ to $r = r_2$ and subtending the angles $\delta\theta$ and $\delta\phi$ as measured at the origin:*



Suppose that a single charge q is placed at the origin, so that S contains no charges but is cut by field lines. The field $\boldsymbol{E}(\boldsymbol{r})$ is radial and so the only contribution to the flux through S comes from the field passing through the sides of S with areas da_1 and

[5]This is the same reason why we can treat a planet of mass M as a point source of the same mass in Newtonian gravity, all courtesy of the $\frac{1}{r^2}$ force law.

da_2 indicated above. If we consider S small enough (and ultimately infinitesimal) then the electric field is approximately constant over each of these surfaces. The flux through the first surface is $-\frac{q}{4\pi\epsilon_0 r_1^2} da_1 = -\frac{q}{4\pi\epsilon_0 r_1^2} r_1^2 \cos\phi \delta\theta\delta\phi = -\frac{q}{4\pi\epsilon_0} \cos\phi\delta\theta\delta\phi$, it is negative because the field lines are entering the surface. While at the second surface the flux is $\frac{q}{4\pi\epsilon_0 r_2^2} da_2 = \frac{q}{4\pi\epsilon_0 r_2^2} r_2^2 \cos\phi\delta\theta\delta\phi = \frac{q}{4\pi\epsilon_0} \cos\phi\delta\theta\delta\phi$, and the total flux through S is zero. Finally suppose that $d\boldsymbol{a}_1$ is at an angle ψ to $\boldsymbol{E}(\boldsymbol{r_1})$, rather than parallel - this will allow us to consider arbitrary shaped bounding surfaces (rather than just spherical surfaces). Then the flux is $-\boldsymbol{E}(\boldsymbol{r_1})\cdot d\boldsymbol{a}_1 = -|\boldsymbol{E}(\boldsymbol{r_1})||d\boldsymbol{a}_1| \cos\psi$. However the area of the infinitesimal surface (extending through $\delta\phi$ and $\delta\theta$) is increased by $\frac{1}{\cos\psi}$ from $r_1^2 \cos\phi\delta\theta\delta\phi$ to $\frac{r_1^2\cos\phi}{\cos\psi}\delta\theta\delta\phi$, so that the flux through the surface is independent of the angle ψ. We may make the same argument for any of the sides of the curving cuboid. Hence, we have confirmed Gauss's law for a closed surface containing no charges, in a non-zero electric field.

We may restate Gauss's law for a charge distribution $\rho(\boldsymbol{r})$ to obtain

$$\oint_S \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{a} = \int_V \frac{\rho(\boldsymbol{r}')}{\epsilon_0} dv'$$

where now $Q_S = \int_V \rho(\boldsymbol{r}')dv'$ and V is the volume bounded by S.

Now Green's theorem for the flux is

$$\oint_S \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{a} = \int_V (\boldsymbol{\nabla} \cdot \boldsymbol{E}(\boldsymbol{r'}))dv'$$

and by comparing the previous two formulae we obtain Gauss's law in differential form:

$$\boldsymbol{\nabla} \cdot \boldsymbol{E}(\boldsymbol{r}) = \frac{\rho(\boldsymbol{r})}{\epsilon_0}.$$

3.3.1 Application of Gauss's Law

The aim of electrostatics is to work out the force on a test charge Q, and since $\boldsymbol{F} = Q\boldsymbol{E}$, this means determining the electric field for a charge configuration. Gauss's law in integral form offers some rapid ways to find $\boldsymbol{E}(\boldsymbol{r})$ for some simplified examples. In these examples which follow each charge distribution will admit a simple choice of surface, called a Gaussian surface, over which the electric field is constant, thereby simplifying the flux integrals. There are three standard Gaussian surfaces which take advantage of rotation and translation symmetries of the surface, and which we will meet in the following examples.

Example 3.1. *(The charged spherical shell) Consider a spherical shell of radius R, centred at the origin, and carrying a uniform distribution of charge of constant charge density ρ. We will compute the electric field for this charge distribution.*

The charge distribution is spherically symmetric and so the electric field is entirely in the radial direction and of equal strength at any point on a surrounding sphere of radius $r > R$. This motivates us to choose the Gaussian surface S for the flux integral to be precisely a sphere of radius $r > R$ centred at the origin. Now from equation (3.4) *we have*

$$\oint_S \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{a} = \frac{Q}{\epsilon_0}$$

where Q is the total charge carried by the spherical shell of radius R, i.e. $Q = 4\pi R^2 \rho$. If we can carry out the flux integral we will be able to find $\boldsymbol{E}(\boldsymbol{r})$. On the sphere the magnitude of $\boldsymbol{E}(\boldsymbol{r})$ is constant due to the spherical symmetry and points in the radial direction, so,

$$\begin{aligned}\oint_S \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{a} &= \int_{\phi=-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{\theta=0}^{2\pi} E(r) r^2 \cos(\phi) d\theta d\phi \\ &= 4\pi r^2 E(r).\end{aligned}$$

Therefore we find that for $r > R$,

$$\boldsymbol{E}(\boldsymbol{r}) = E(r)\, \boldsymbol{e_r} = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^2} \boldsymbol{e_r} = \frac{1}{\epsilon_0} \frac{R^2 \rho}{r^2} \boldsymbol{e_r}.$$

For $r < R$ the charge distribution lies outside the surface S hence for $r < R$ we have

$$\oint_S \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{a} = 0$$

and so for $r < R$

$$\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{0}.$$

Altogether we have computed that

$$\boldsymbol{E}(\boldsymbol{r}) = \begin{cases} \frac{1}{\epsilon_0} \frac{R^2 \rho}{r^2} \boldsymbol{e_r} & \text{for } r > R \\ \boldsymbol{0} & \text{for } r < R \end{cases}$$

We see that the electric field is discontinuous across the spherical shell of charge. We also see that outside the shell of charge ($r > R$) the electric field is the same as if we had been working with a point charge with $q = 4\pi R^2 \rho$ located at the origin.

Example 3.2. *(An infinitely long charged line.) Consider an infinitely long, thin wire aligned along the z-axis and carrying a linear charge density of λ. Compute the electric field for this charge distribution.*

The charge distribution has cylindrical symmetry, so we will adopt cylindrical coordinates. Our Gaussian surface will be a cylinder of radius ρ about the z-axis and of length L. Due to the symmetry of the charge distribution, the electric field will point in the (cylindrical) radial direction $\boldsymbol{e_\rho}$, and will be of constant magnitude on the cylindrical Gaussian surface. On the cylinder we have $d\boldsymbol{a} = \rho d\theta dz\, \boldsymbol{e_\rho}$. Hence from equation (3.4) we have

$$\begin{aligned}\frac{Q}{\epsilon_0} &= \oint_S \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{a} \\ &= \int_{\theta=0}^{2\pi} \int_0^L (E(\rho)\, \boldsymbol{e_\rho}) \cdot (\rho d\theta dz\, \boldsymbol{e_\rho}) \\ &= E(\rho)\rho \int_{\theta=0}^{2\pi} d\theta \int_0^L dz \\ &= E(\rho)\rho 2\pi L\end{aligned}$$

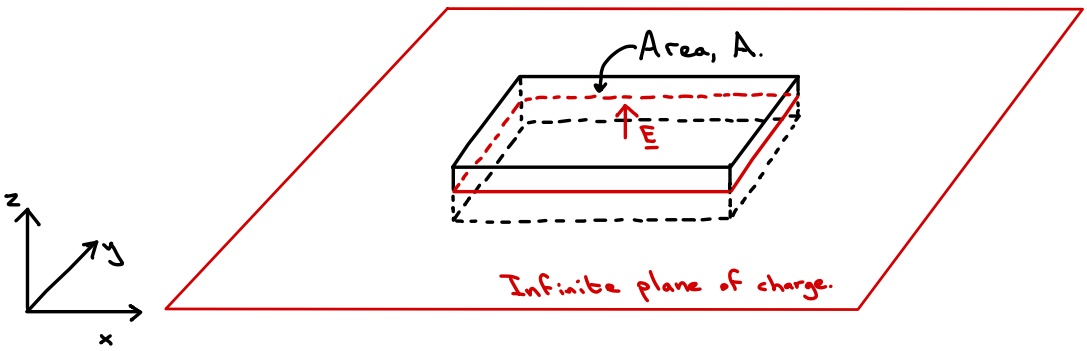
where $Q = L\lambda$, so, $E(\rho) = \frac{\lambda}{2\pi\rho\epsilon_0}$ hence

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{\lambda}{2\pi\rho\epsilon_0}\, \boldsymbol{e_\rho}.$$

for $\boldsymbol{r} = \rho\, \boldsymbol{e_\rho} + \theta\, \boldsymbol{e_\theta} + z\, \boldsymbol{e_z}$ where $\theta \in [0, 2\pi)$ and $z \in [0, L]$ (i.e. $\boldsymbol{r}$ lies on the cylindrical Gaussian surface).

Example 3.3. *(The charged infinite plane) An infinite plane carries a uniform surface charge density σ. Find the electric field due to this charge configuration.*

The translational symmetry of the plane implies that the electric field will be perpendicular to the plane. Hence we should choose a matchbox shaped Gaussian surface for S. Let the infinite plane coincide with the xy-plane. Now consider a matchbox (a cuboid) oriented half below and half above the charged plane:



The sides of area A of the surface are parallel to the plane, and these are the only sides of the matchbox through which there is any flux since $\boldsymbol{E}(\boldsymbol{r}) = \pm E(r)\,\boldsymbol{e_z}$*, we denote these sides* S_1 *for the top side and* S_2 *for the lower side. Hence from Gauss's law in equation* (3.4) *we have,*

$$\begin{aligned}\frac{Q}{\epsilon_0} &= \oint_S \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{a} \\ &= \int_{S_1} (E(r)\,\boldsymbol{e_z}) \cdot (dxdy\;\boldsymbol{e_z}) + \int_{S_2} (-E(r)\,\boldsymbol{e_z}) \cdot (-dxdy\;\boldsymbol{e_z}) \\ &= 2AE(r)\end{aligned}$$

where we have contributions to the flux from the top side and the bottom side of the matchbox, but no other. Now $Q = A\sigma$*, so we find*

$$\boldsymbol{E}(\boldsymbol{r}) = \begin{cases} \frac{\sigma}{2\epsilon_0}\,\boldsymbol{e_z} & \text{for } z > 0 \\ -\frac{\sigma}{2\epsilon_0}\,\boldsymbol{e_z} & \text{for } z < 0 \end{cases}$$

where $\boldsymbol{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ *lies on the Gaussian matchbox.*

This is an amazing result: the electric field does not diminish no matter how far one measures it from the plane. This is a consequence of the electric plane being infinite in extent and the problem being scale invariant: a calculation for a point a distance r *could be rescaled to be similar to a calculation of the field at* $R > r$*, even before any computation is done we can see that the problems are identical as the infinite plane of charge makes it impossible to measure scale, hence any result should be independent of* r*. Another way to think about this result is that the field lines from the infinite charged plane keep the same density all the way off to infinity, therefore the flux does not depend on distance from the charged plane.*

In summary, the electric field falls off as $\frac{1}{r^2}$ for a point source, or for a uniformly charged spherical shell; it falls off as $\frac{1}{r}$ for a charged infinite line; and it stays constant for the charged infinite plane.

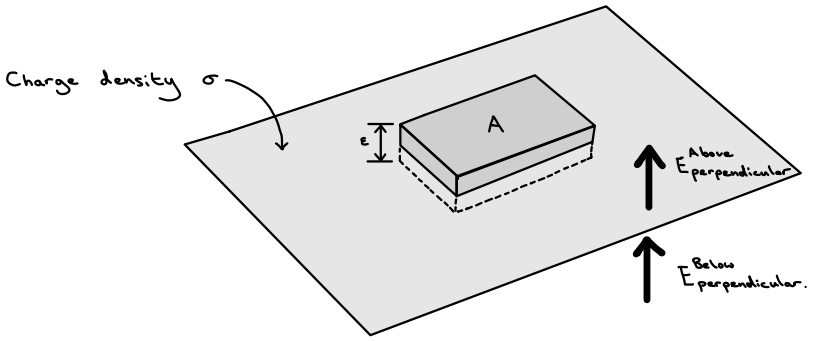
In the examples above we could use symmetries of the charge distribution to pick out a Gaussian surface on which $\boldsymbol{E}(\boldsymbol{r})$ is constant to compute the flux. However the method above is only useful when we have spherical, cylindrical or planar symmetry. What can we do to find $\boldsymbol{E}(\boldsymbol{r})$ in the absence of these symmetries? One idea is that we can use the superposition of electric fields to find more complicated electric fields, which feature combinations of spherical, cylindrical and planar charge distributions: $\boldsymbol{E}(\boldsymbol{r}) = \sum_{i=1}^{n} \boldsymbol{E_i}(\boldsymbol{r})$. We will consider some examples of this in the problem sets. In many cases we can also sketch field lines fairly quickly to get a rough idea of $\boldsymbol{E}(\boldsymbol{r})$, but to find $\boldsymbol{E}(\boldsymbol{r})$ precisely in most circumstances we will use other methods, beginning with concept of the electric potential.

3.3.2 Changes in the Electric Field Across a Charged Surface

In the last example, we discovered that the electric field is discontinuous across the charged infinite plane. If we write $\boldsymbol{E}^{Above} = \frac{\sigma}{2\epsilon_0}\boldsymbol{e_z}$ for the field above ($z > 0$) the charged plane and $\boldsymbol{E}^{Below} = -\frac{\sigma}{2\epsilon_0}\boldsymbol{e_z}$ for the field below ($z < 0$) the charged plane then we find

$$\boldsymbol{E}^{Above} - \boldsymbol{E}^{Below} = \frac{\sigma}{\epsilon_0}\boldsymbol{e_z}.$$

This discontinuity in the electric field normal to a charged surface is a result applicable to the component of an electric field normal to any charged surface. The difference in the general case, is that there may be other sources of the electric field besides the charge surface, however the only change this makes to the computation is that we can no longer argue that $\boldsymbol{E}(\boldsymbol{r})$ is perpendicular to the surface, because the general surface does not possess translational symmetry. However this does not alter much about the computation. To see this we need only repeat the computation in the example above, but allow the surface to have an arbitrary (but continuous) shape and orientation. We may imagine a Gaussian surface which is a matchbox of infinitesimal height ϵ surrounding a patch of area A on the charged surface:



In the limit where ϵ tends to zero, there is zero contribution to the flux of $\boldsymbol{E}(\boldsymbol{r})$ through the matchbox from its sides (recall that in this more general picture we are not assuming that $\boldsymbol{E}(\boldsymbol{r})$ is perpendicular to the surface). The only contribution to the flux is from the top and bottom face hence we find

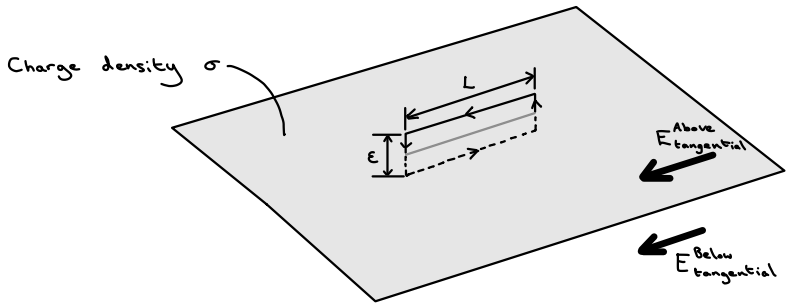
$$\oint_S \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{a} = E^{Above}_{perpendicular} A - E^{Below}_{perpendicular} A = \frac{\sigma A}{\epsilon_0}$$

hence

$$E^{Above}_{perpendicular} - E^{Below}_{perpendicular} = \frac{\sigma}{\epsilon_0}. \tag{3.5}$$

Therefore the normal component of $\boldsymbol{E}(\boldsymbol{r})$ is discontinuous at any charged surface and jumps by an amount $\frac{\sigma}{\epsilon_0}$ where σ is the charge density. This is a very useful result, and it will help us think about electric fields in diverse settings.

The tangential component of $\boldsymbol{E}(\boldsymbol{r})$ is always continuous across a charged surface. To see this we will use a result that we will deduce in the next section namely that $\oint_P \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{l} = 0$. Consider the closed path that runs around the rectangle of sides L and ϵ, where the edges of length L are parallel to the surface (i.e. for a sufficiently small L) and the edges of infinitesimal length ϵ are perpendicular to the surface as depicted here:



Now consider the line integral of $\boldsymbol{E}(\boldsymbol{r})$ around the closed path P around the edges of the rectangle in the limit where ϵ goes to zero:

$$\lim_{\epsilon \to 0} \Big[\oint_P \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{l} \Big] = E^{Above}_{tangential} L - E^{Below}_{tangential} L = 0$$

where we have used $\oint_P \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{l} = 0$. Hence, at any charged surface,

$$E^{Above}_{tangential} = E^{Below}_{tangential}.$$

It remains for us to prove that $\oint_P \boldsymbol{E}(\boldsymbol{r}) \cdot d\boldsymbol{l} = 0$ and for this we will need to introduce the electric potential.

3.4 The Electric Potential

By computing $\boldsymbol{\nabla} \cdot \boldsymbol{E}$ we have deduced Gauss's law, so we may be wondering about what may be inferred from $\boldsymbol{\nabla} \times \boldsymbol{E}$. Consider the electric field of a single point charge q located at the origin

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \boldsymbol{e_r}.$$

This is the product of a scalar function $f(r) = \frac{1}{4\pi\epsilon_0}\frac{1}{r^2}$ and a unit radial vector[6] $\boldsymbol{v} = \boldsymbol{e_r}$, so to take the curl we use the product law given in equation (2.9) and the del operator $\boldsymbol{\nabla}$ in spherical

[6]We note that although $\boldsymbol{e_r}$ has constant length of one, its direction varies. As a result the vector $\boldsymbol{e_r}$ has non-zero derivatives $\frac{\partial}{\partial\theta}(\boldsymbol{e_r}) = \cos\phi\, \boldsymbol{e_\theta}$, $\frac{\partial}{\partial\phi}(\boldsymbol{e_r}) = \boldsymbol{e_\phi}$.

coordinates, as given in equation (C.6) to find

$$\boldsymbol{\nabla}\times\boldsymbol{E}=\boldsymbol{\nabla}(\frac{1}{4\pi\epsilon_0}\frac{q}{r^2})\times\boldsymbol{e_r}+\frac{1}{4\pi\epsilon_0}\frac{q}{r^2}(\boldsymbol{\nabla}\times\boldsymbol{e_r})$$

Working in spherical coordinates[7] we see that $\boldsymbol{e_r}$ is depends upon θ and ϕ (see appendix C) so that $\frac{\partial}{\partial\theta}(\boldsymbol{e_r})=\cos\phi\,\boldsymbol{e_\theta}$ and $\frac{\partial}{\partial\phi}(\boldsymbol{e_r})=\boldsymbol{e_\phi}$.

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{e_r}&=\Big(\boldsymbol{e_r}\frac{\partial}{\partial r}+\boldsymbol{e_\theta}\frac{1}{r\cos\theta}\frac{\partial}{\partial\theta}+\boldsymbol{e_\phi}\frac{1}{r}\frac{\partial}{\partial\phi}\Big)\times\boldsymbol{e_r}\\&=(\boldsymbol{e_\theta}\frac{1}{r\cos\theta})\times(\cos\phi\,\boldsymbol{e_\theta})+(\boldsymbol{e_\phi}\frac{1}{r})\times\boldsymbol{e_\phi}\\&=\boldsymbol{0}.\end{aligned}$$

So that we find[8],

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{E}&=\boldsymbol{\nabla}(\frac{1}{4\pi\epsilon_0}\frac{q}{r^2})\times\boldsymbol{e_r}\\&=\frac{1}{4\pi\epsilon_0}\frac{\partial}{\partial r}\Big(\frac{q}{r^2}\Big)\boldsymbol{e_r}\times\boldsymbol{e_r}\\&=\boldsymbol{0}.\end{aligned}$$

Now since we can produce the electric field of any static arrangement of charges by vector addition of the electric fields for each charge (or by integration for a distribution) then we find that in general, for any charge distribution,

$$\boldsymbol{\nabla}\times\boldsymbol{E}=\boldsymbol{\nabla}\times[\boldsymbol{E_1}+\boldsymbol{E_2}+\boldsymbol{E_3}+\ldots]=\boldsymbol{\nabla}\times\boldsymbol{E_1}+\boldsymbol{\nabla}\times\boldsymbol{E_2}+\boldsymbol{\nabla}\times\boldsymbol{E_3}+\ldots=\boldsymbol{0}.$$

We may also have deduced that the curl of $\boldsymbol{E}(\boldsymbol{r})$ for a point charge is zero by observing that

$$\boldsymbol{E}(\boldsymbol{r})=\frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\boldsymbol{e_r}=-\boldsymbol{\nabla}\Big(\frac{1}{4\pi\epsilon_0}\frac{q}{r}\Big)\equiv-\boldsymbol{\nabla}\Phi$$

where Φ is a scalar function which is the electric potential (for the point charge in this case) and hence $\boldsymbol{\nabla}\times\boldsymbol{E}=\boldsymbol{\nabla}\times(-\boldsymbol{\nabla}\Phi)=\boldsymbol{0}$, since the curl of a gradient is the zero vector. For a

[7] Of course we can also quickly use Cartesian coordinates to confirm that $\boldsymbol{\nabla}\times\boldsymbol{e_r}=\boldsymbol{0}$, as $\boldsymbol{e_r}=\frac{\boldsymbol{r}}{r}=\frac{1}{\sqrt{x^2+y^2+z^2}}\begin{pmatrix}x\\y\\z\end{pmatrix}$, hence $\boldsymbol{\nabla}\times\boldsymbol{e_r}=(-\frac{zy}{r^3}+\frac{yz}{r^3})\boldsymbol{e_x}+(-\frac{xz}{r^3}+\frac{zx}{r^3})\boldsymbol{e_y}+(-\frac{yx}{r^3}+\frac{xy}{r^3})\boldsymbol{e_z}=\boldsymbol{0}$. But it is a good idea to practise using spherical and cylindrical coordinates, and in particular to recall that the derivatives always act before the vector product when computing curl and div.

[8] In Cartesian coordinates we would compute $\boldsymbol{\nabla}\frac{1}{r^2}=-2\frac{x}{r^4}\boldsymbol{e_x}-2\frac{y}{r^4}\boldsymbol{e_y}-2\frac{z}{r^4}\boldsymbol{e_z}=-\frac{2}{r^4}\boldsymbol{r}=-\frac{2}{r^3}\boldsymbol{e_r}$.

general charge distribution we observe that

$$\begin{aligned}\Big(\boldsymbol{\nabla}\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\Big) &= -\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|^3}\Big((x-x')\,\boldsymbol{e_x}+(y-y')\,\boldsymbol{e_y}+(z-z')\,\boldsymbol{e_z}\Big)\\ &= -\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|^3}(\boldsymbol{r}-\boldsymbol{r'})\\ &= -\frac{1}{s^2}\boldsymbol{e_s}\end{aligned}$$

and so

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r'})}{s^2}\,\boldsymbol{e_s}\,dv' = -\boldsymbol{\nabla}\Big(\frac{1}{4\pi\epsilon_0}\int\frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}dv'\Big) = -\boldsymbol{\nabla}\Phi$$

implying that $\boldsymbol{\nabla}\times\boldsymbol{E}=\boldsymbol{0}$.

The scalar function Φ satisfying $\boldsymbol{E}(\boldsymbol{r}) = -\boldsymbol{\nabla}\Phi$ is called the electric potential, and for a charge distribution $\rho(\boldsymbol{r'})$ is given by

$$\Phi = \frac{1}{4\pi\epsilon_0}\int\frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}dv'. \tag{3.6}$$

Comment(s).

- *Since $\boldsymbol{E}=-\boldsymbol{\nabla}\Phi$ then the electric field at any point in $\mathbb{R}^3$ gives a vector which points in the direction of fastest decrease in the potential Φ.*

- *A surface upon which the potential is constant is called an equipotential surface. Consider a vector field $\boldsymbol{t}_S$ which is tangent to the equipotential surface, S, i.e. for any motion along S the potential remains constant. Writing this in terms of the directional derivative in the direction $\boldsymbol{t}_S$ gives $(\boldsymbol{t}_S\cdot\boldsymbol{\nabla})\Phi=0$. However this is the same statement as $\boldsymbol{t}_S\cdot(\boldsymbol{\nabla}\Phi) = \boldsymbol{t}_S\cdot(-\boldsymbol{E}) = 0$. Therefore the electric field lines are perpendicular to the equipotential surfaces.*

- *The electric field being a vector field is defined by three scalar functions, while the potential is a single scalar function from which the electric field may be derived. In short the electric potential may be simpler to work with, but contains the same information as the electric field.*

- *Since the electric field is (minus) the gradient of the electric potential then the path integral*

from $\boldsymbol{a}$ *to* $\boldsymbol{b}$ *is path-independent:*

$$\begin{aligned}\int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{E}\cdot d\boldsymbol{l} &= -\int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{\nabla}\Phi\cdot d\boldsymbol{l}\\ &= -\int_{\boldsymbol{a}}^{\boldsymbol{b}} \Big(\frac{\partial\Phi}{\partial x}dx+\frac{\partial\Phi}{\partial y}dy+\frac{\partial\Phi}{\partial z}dz\Big)\\ &= -\int_{\boldsymbol{a}}^{\boldsymbol{b}} d\Phi\\ &= -\Phi(\boldsymbol{b})+\Phi(\boldsymbol{a})\end{aligned}$$

and depends on only the values of the potential at the beginning and end of the path. Consequently the path integral of $\boldsymbol{E}(\boldsymbol{r})$ *around a closed loop is zero. This is something we could have deduced by applying Stokes' theorem to* $\boldsymbol{E}(\boldsymbol{r})$*:*

$$\oint_P \boldsymbol{E}\cdot d\boldsymbol{l} = \int_S (\boldsymbol{\nabla}\times\boldsymbol{E})\cdot d\boldsymbol{a} = 0.$$

- *The potential is defined by* $\boldsymbol{E}(\boldsymbol{r}) = -\boldsymbol{\nabla}\Phi$ *hence it is only defined up to a constant* C*, as* $-\boldsymbol{\nabla}(\Phi+C) = -\boldsymbol{\nabla}\Phi = \boldsymbol{E}(\boldsymbol{r})$*. The force on a test charge* $\boldsymbol{F} = Q\boldsymbol{E}$ *is measurable and takes the same value no matter which constant we choose for* C*. The potential is not measurable (although the potential difference is) and we can choose where we would like to set it to zero. In most circumstances*[9] *it is sensible to set the potential to zero at infinity, as in most cases, the electric field also diminishes to zero. In which case* $\int_{\boldsymbol{a}}^{\infty}\boldsymbol{E}\cdot d\boldsymbol{l} = \Phi(\boldsymbol{a})$*. If we denote by* $\boldsymbol{O}$ *the point where the potential is set to zero then we may write*[10]

$$\Phi(\boldsymbol{r}) = -\int_{\boldsymbol{O}}^{\boldsymbol{r}} \boldsymbol{E}\cdot d\boldsymbol{l}.$$

- *The electric potential obeys the superposition principle as if* $\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{E_1}+\boldsymbol{E_2}+\ldots$ *then* $\boldsymbol{\nabla}\Phi = \boldsymbol{\nabla}\Phi_1+\boldsymbol{\nabla}\Phi_2+\ldots = \boldsymbol{\nabla}(\Phi_1+\Phi_2+\ldots)$*, hence* $\Phi = \Phi_1+\Phi_2+\ldots$*.*

- *The potential is measured in volts, which we may deduce from Coulomb's law is equivalent to Newton-metres*[11] *per Coulomb.*

Example 3.4. *Sketch a selection of equipotential surfaces together with the electric field lines for*

[9] The infinitely charged plane is an example in which the electric field is constant and does not fall off towards infinity, in such cases one must choose an alternative zero for the potential.

[10] It's customary to keep the minus sign and write the limits in this order, as it signifies the idea of bringing in a test charge from infinity (or $\mathbf{O}$) to $\boldsymbol{r}$.

[11] A Newton-metre, the unit of work and hence energy, is also called a joule.

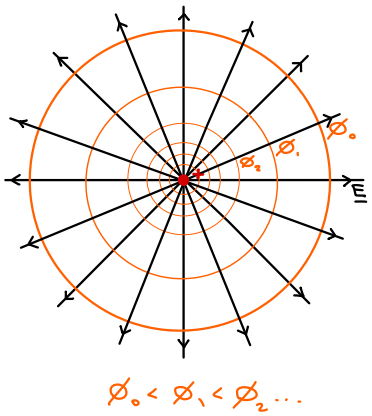
(a) a point source charge (sketch a diagram in which the source charge is positive and another diagram in which it is negative),

(b) a pair of source charges where one has charge q while the other has charge $-q$, and

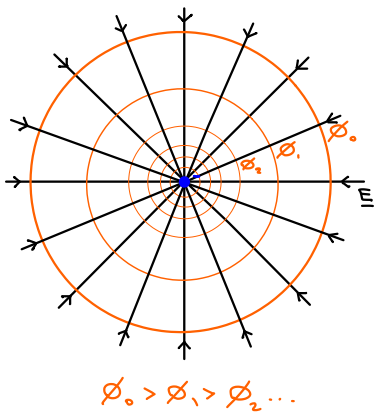
(c) a pair of source charges where both have charge q.

For each of the sketches we will use the observation that equipotential surfaces (which will be shown as equipotential lines in our two-dimensional cross-sections) are at right angles to the electric field.

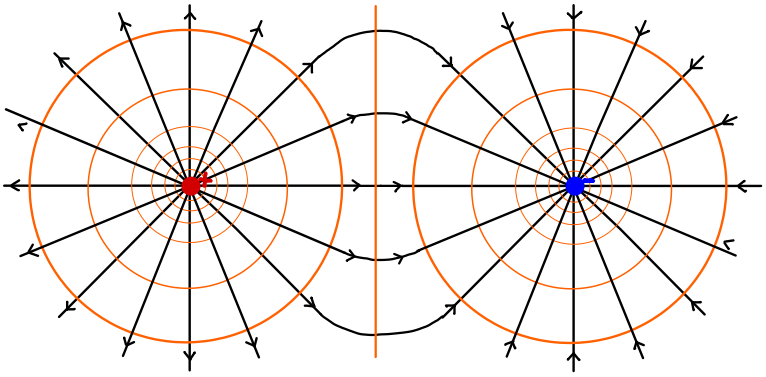
(a) The positive charge is a maximum of the potential:



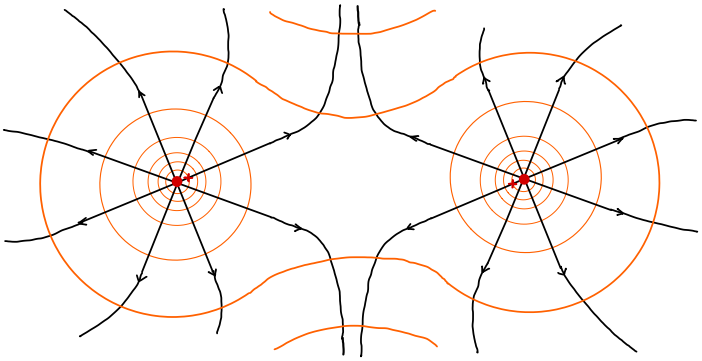
The negative charge is a minimum of the potential:



(b)



(c)



Example 3.5. *Find the potential inside and outside a ball of radius R of constant charge density ρ.*

We may use use Gauss's law to quickly determine $\boldsymbol{E}(\boldsymbol{r})$ due to the spherical symmetry. Outside the ball, on a sphere of radius $r > R$ we have

$$E(r)(4\pi r^2) = \oint_S \boldsymbol{E} \cdot d\boldsymbol{a} = \frac{Q}{\epsilon_0} = \frac{4\pi R^3 \rho}{3\epsilon_0}$$

hence

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{R^3 \rho}{3\epsilon_0 r^2}\, \boldsymbol{e_r} \qquad \textit{for} \quad r > R.$$

While inside the sphere we may do the same computation for a sphere of radius $r < R$, which will diminish the total charge inside the surface of the integral:

$$E(r)(4\pi r^2) = \oint_S \boldsymbol{E} \cdot d\boldsymbol{a} = \frac{Q}{\epsilon_0} = \frac{4\pi r^3 \rho}{3\epsilon_0}$$

hence

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{r \rho}{3\epsilon_0}\, \boldsymbol{e_r} \qquad \textit{for} \quad r < R.$$

Then using $\Phi(\boldsymbol{r}) = -\int_\infty^{\boldsymbol{r}} \boldsymbol{E}\cdot d\boldsymbol{l}$*, where we have defined that* Φ *is zero at infinity and for a radial path from infinity to* $\boldsymbol{r}$*, we find for* $r > R$*:*

$$\begin{aligned}\Phi(r) &= -\int_\infty^r \left(\frac{R^3\rho}{3\epsilon_0 r^2}\,\boldsymbol{e_r}\right)\cdot(dr\,\boldsymbol{e_r}) \\ &= -\frac{R^3\rho}{3\epsilon_0}\int_\infty^r \frac{1}{r^2}dr \\ &= \frac{R^3\rho}{3\epsilon_0 r}\end{aligned}$$

which is as expected. Inside the ball, for $r < R$ *we have also to consider the contribution to the potential from outside the ball, as it is defined by an integral from* ∞ *to the radius* r*, hence, we split the integral into two parts:*

$$\begin{aligned}\Phi(r) &= -\int_R^r \left(\frac{r\rho}{3\epsilon_0}\,\boldsymbol{e_r}\right)\cdot(dr\,\boldsymbol{e_r}) - \int_\infty^R \left(\frac{R^3\rho}{3\epsilon_0 r^2}\,\boldsymbol{e_r}\right)\cdot(dr\,\boldsymbol{e_r}) \\ &= -\int_R^r \frac{r\rho}{3\epsilon_0}dr + \frac{R^3\rho}{3\epsilon_0 R} \\ &= -\frac{r^2\rho}{6\epsilon_0} + \frac{R^2\rho}{2\epsilon_0}.\end{aligned}$$

Notice that the electric potential inside the ball depends upon the electric field both inside and outside the charged ball.

3.5 Work and Energy

Electrostatics involves investigating the force felt by a test charge Q due to the static arrangement of source charges. Since the test charge feels a force we might be interested in the work done to bring the test charge from infinity to $\boldsymbol{r}$ against the field - this would give us some measure of the energy of the configuration. However the situation is a precarious one, since we are working with electrostatics, our assumptions have been to disregard motion in the system, and in thinking about the work done to move a charge against the field we will be imposing the unrealistic condition that the source charges remain static. Nevertheless the result will be interesting and will give us more insight into the electric potential. Now the Coulomb force $\boldsymbol{F}$ acts on the test charge Q, so the force needed to act on Q to balance out the Coulomb force is

$-\boldsymbol{F}$. We may compute the work done to bring Q to $\boldsymbol{r}$ from infinity as follows

$$\begin{aligned} W &= -\int_{\infty}^{\boldsymbol{r}} \boldsymbol{F}\cdot d\boldsymbol{l} \\ &= -Q\int_{\infty}^{\boldsymbol{r}} \boldsymbol{E}\cdot d\boldsymbol{l} \\ &= Q\int_{\infty}^{\boldsymbol{r}} \boldsymbol{\nabla}\Phi\cdot d\boldsymbol{l} \\ &= Q\int_{\infty}^{\boldsymbol{r}} d\Phi \\ &= Q\Phi(\boldsymbol{r}). \end{aligned}$$

So that the work per unit charge $\frac{W}{Q}$ is equal to the potential $\Phi(\boldsymbol{r})$. The potential is really the potential energy per unit charge.

Furthermore if we had considered moving a test charge from $\boldsymbol{a}$ to $\boldsymbol{b}$ we would have had to do the work per unit charge:

$$\frac{W}{Q} = \Phi(\boldsymbol{b}) - \Phi(\boldsymbol{a})$$

so that the potential difference between two points $\boldsymbol{a}$ and $\boldsymbol{b}$ is equal to the work per unit charge needed to move Q from $\boldsymbol{a}$ to $\boldsymbol{b}$.

This idea, despite being somewhat artificial, allows us to compute the potential energy of a particular configuration of charges. We may simply compute the work needed to bring each charge q_i from infinity to its position in the charge distribution, as we gradually build up the charge configuration. Suppose that we have n source charges q_1, q_2,... q_n at positions $\boldsymbol{r_1'}$, $\boldsymbol{r_2'}$, ... $\boldsymbol{r_n'}$ and we imagine bringing each charge to its position from infinity. The first charge q_1 requires no work to be done to put it in position, as there is no background electric field: $W_1 = 0$. However when we bring in the second charge q_2 it will feel the force due to the electric field of q_1, whose potential is $\Phi_1(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\frac{q_1}{|\boldsymbol{r}-\boldsymbol{r_1'}|}$, hence the work to bring the second particle into position is

$$W_2 = q_2\Phi_1(\boldsymbol{r_2'}) = \frac{1}{4\pi\epsilon_0}\frac{q_1q_2}{|\boldsymbol{r_2'}-\boldsymbol{r_1'}|} = \frac{1}{4\pi\epsilon_0}\frac{q_1q_2}{s_{12}}$$

where $s_{12} = |\boldsymbol{r_2'}-\boldsymbol{r_1'}|$. The potential of the two charges is found by summing the potentials of the two isolated charges:

$$\Phi_2(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\Big(\frac{q_1}{|\boldsymbol{r}-\boldsymbol{r_1'}|} + \frac{q_2}{|\boldsymbol{r}-\boldsymbol{r_2'}|}\Big)$$

so the work to bring the third charge in from infinity is

$$W_3 = q_3\Phi_2(\boldsymbol{r_3'}) = \frac{1}{4\pi\epsilon_0}q_3\Big(\frac{q_1}{s_{13}} + \frac{q_2}{s_{23}}\Big).$$

We may now generalise and deduce that the work required to bring the k'th charge into the configuration is

$$W_k = q_k \Phi_{k-1}(\boldsymbol{r'_k}) = \frac{1}{4\pi\epsilon_0} q_k \Big(\frac{q_1}{s_{1k}} + \frac{q_2}{s_{2k}} + \ldots + \frac{q_{k-1}}{s_{(k-1)k}} \Big) = \frac{1}{4\pi\epsilon_0} q_k \sum_{i=1}^{k-1} \frac{q_i}{s_{ik}}$$

with the understanding that $W_1 = 0$. Then the total work to construct the configuration of n charges is

$$\begin{aligned} W &= \sum_{j=1}^{n} W_j \\ &= \frac{1}{4\pi\epsilon_0} \sum_{j=1}^{n} q_j \sum_{i<j} \frac{q_i}{s_{ij}} \\ &= \frac{1}{4\pi\epsilon_0} \sum_{i=1}^{n} \sum_{j>i} \frac{q_j q_i}{s_{ij}}. \end{aligned}$$

Hence we may write

$$2W = \frac{1}{4\pi\epsilon_0} \Big(\sum_{j=1}^{n} \sum_{i<j} \frac{q_j q_i}{s_{ij}} + \sum_{i=1}^{n} \sum_{j>i} \frac{q_j q_i}{s_{ij}} \Big) = \frac{1}{4\pi\epsilon_0} \sum_{j=1}^{n} \sum_{i \neq j} \frac{q_i q_j}{s_{ij}}.$$

Now $\frac{1}{4\pi\epsilon_0} \sum_{i \neq j} \frac{q_i}{s_{ij}} = \Phi(\boldsymbol{r'_j})$ is the potential at $\boldsymbol{r'_j}$ due to all the other charges (apart from q_j). Therefore

$$W = \frac{1}{2} \sum_{i=1}^{n} q_i \Phi(\boldsymbol{r'_i})$$

is a measure of the potential energy of the static system, all that it is missing is any energy required to hold still the configuration of charges as the charges are brought in from infinity.

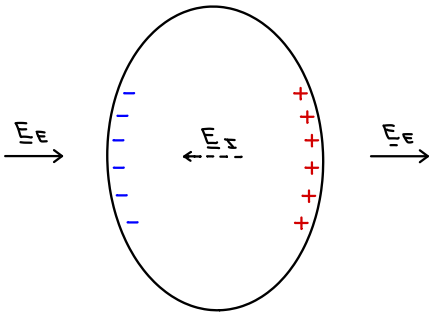
3.6 Conductors

A conductor is a material in which some charges are free to move, but yet constrained to remain within the boundaries of the material. For example metal is a conductor, in which electrons may move effectively freely within the material.

Problems in electrostatics concern static charges, so one might think that conductors are not discussed in electrostatics. In practise the free electrons in a conductor, when subjected to an electrostatic force, rapidly reach an equilibrium (for a metal, equilibrium is typically reached in tens of hundreds of femtoseconds 10^{-15} seconds). Consequently the kinds of distribution of charges found in a conductor may be investigated.

Comment(s).

- $\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{0}$ *inside a conductor*[12]. *When placed in an external electric field* $\boldsymbol{E}_E$ *the free charges in a conductor move subject to the forces associated with* $\boldsymbol{E}_E$*, and they are free to move until they reach the surface of the conductor. For a metal, it is the negatively charged electrons which move - they move in the opposite direction to the arrowheads on the field lines of* $\boldsymbol{E}_E$ *(as the field lines show the direction a positive test charge would move in). The electrons leave behind positively charged atoms, so that initially an induced electric field* $\boldsymbol{E}_I$ *is set up in the conductor whose field lines oppose those of* $\boldsymbol{E}_E$*.*



The induced field grows until it balances the external field inside the conductor and equilibrium is reached. The process of the electric field being nullified within a conductor is almost instantaneous.

- *Due to the strength of the electromagnetic forces, which very rapidly cancels out unbalanced charge, conductors may be expected to be neutrally charged. Although they contain negative and positive charges, they are frequently in balance over the conductor. Inside a conductor, Gauss's law tells us that* $\oint_S \boldsymbol{E} \cdot d\boldsymbol{a} = 0 = \frac{Q}{\epsilon_0}$*, and so* $Q = 0$ *inside a conductor as* $\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{0}$*. If* Q *arises from a charge distribution then* $\rho = 0$*. However suppose that there was some net charge in a conductor, then it would move until it reaches the surface of the conductor, where it would be trapped. Any net charge lies on the surface of a conductor.*

- *The potential inside a conductor is constant:*

$$\begin{aligned}\Phi(\boldsymbol{r}) &= -\int_{\infty}^{\boldsymbol{r}} \boldsymbol{E} \cdot d\boldsymbol{l} \\ &= -\int_{\infty}^{\boldsymbol{r_s}} \boldsymbol{E} \cdot d\boldsymbol{l} - \int_{\boldsymbol{r_s}}^{\boldsymbol{r}} \boldsymbol{E} \cdot d\boldsymbol{l} \\ &= \Phi(\boldsymbol{r_s})\end{aligned}$$

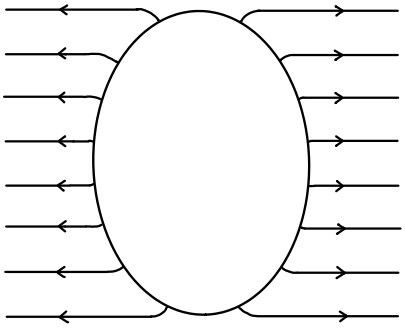
[12] Or at least within hundreds of femtoseconds of being placed in an electric field, $\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{0}$.

where $\boldsymbol{r_s}$ is a radial vector on the surface of the conductor, which at first seems to depend on the path P taken in the line integral. However any point $\boldsymbol{r_S}$ on the surface of the conductor will do as the line integral of the electric field is path-independent. We may convince ourselves further as for any two points $\boldsymbol{a}$ and $\boldsymbol{b}$ in the conductor (or on its surface) then

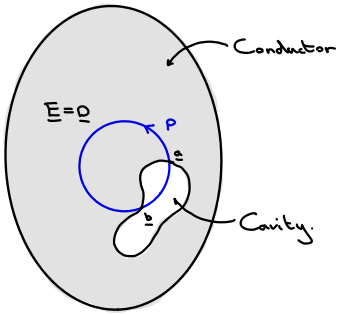
$$0 = \int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{E} \cdot d\boldsymbol{l} = \Phi(\boldsymbol{a}) - \Phi(\boldsymbol{b})$$

hence all points on or in the conductor are at the same potential: it is an equipotential line, surface or volume.

- *As $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$, then an electric field is perpendicular to any equipotential surface. Therefore the field lines must lie perpendicular to the surface of a conductor. If there were any field lines tangential to the surface of the conductor they would induce an electric field in the surface of the conductor that rapidly cancels out the tangential component of the external field.*



- *Within an empty cavity inside a conductor $\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{0}$. Since a cavity inside a conductor introduces a new internal surface, it is conceivable that charge may accumulate on this inner surface, just as it does on the outer surface of a conductor, such a situation would lead to electric field lines starting on positive charge on the cavity's boundary and ending on negative charge somewhere else on the cavity's boundary.*



Consider the line integral around any closed path P, contained within the conductor and passing through the cavity:

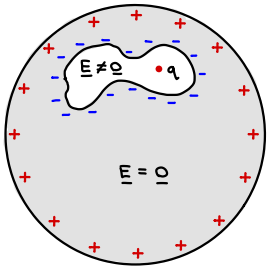
$$\oint_P \boldsymbol{E} \cdot d\boldsymbol{l} = \int_{\boldsymbol{b}}^{\boldsymbol{a}} \boldsymbol{E} \cdot d\boldsymbol{l} = 0$$

as $\oint \boldsymbol{E} \cdot d\boldsymbol{l} = 0$. Note that we have used the fact that $\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{0}$ inside the conductor. This argument holds for any closed path P so we deduce that $\int_{\boldsymbol{b}}^{\boldsymbol{a}} \boldsymbol{E} \cdot d\boldsymbol{l} = 0$ for any $\boldsymbol{a}$ and vecb on the boundary of the cavity and for any path through the cavity. In particular if we take a path that runs along a hypothetical electric field line in the cavity then $\int_{\boldsymbol{b}}^{\boldsymbol{a}} \boldsymbol{E} \cdot d\boldsymbol{l} \geq 0$ hence it must be that $\boldsymbol{E} = \boldsymbol{0}$ within the cavity (as well as within the conductor). Consequently, using Gauss's law for a surface S within the conductor that includes the wall of the cavity in its interior gives

$$\frac{Q_S}{\epsilon_0} = \oint \boldsymbol{E} \cdot d\boldsymbol{a} = 0$$

as $\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{0}$ on the surface. Hence there is no net charge accumulating on the surface of the interior cavity. The vanishing of $\boldsymbol{E}(\boldsymbol{r})$ within a cavity surrounded by a conductor, effectively shields the region from any external electric field. This is the idea behind the Faraday cage.

- *If a charge q is placed within a cavity inside a conductor then there is a non-zero electric field associated with the charge inside the cavity. The field in the surrounding conductor is again zero: for the same reasons that $\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{0}$ inside a conductor: charge redistributes on the surface of the cavity so that the induced field cancels out the charges electric field in the conductor. E.g. for a positive source charge placed within a cavity inside a conductor, negative charge will accumulate on the cavity's surface, and the positive charge of the source charge will distribute over the outer surface of the conductor:*



This demonstrates a practical way in which point charge sources can be converted into, effectively continuous, distributed charge sources. Notice that the external field then depends only on the shape of the conductor, and not the cavity. So that a point charge in an irregularly shaped cavity within a spherical conductor will have the external field of a

charged spherical shell, which, outside the sphere, is identical to that of a single charge located at the centre of the sphere. The shape of the cavity plays no role.

4. Potentials

Our work, so far, has reduced electrostatics to the problem of solving the following pair of equations to find $\boldsymbol{E}(\boldsymbol{r})$:

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}, \quad \text{and} \tag{4.1}$$

$$\boldsymbol{\nabla} \times \boldsymbol{E} = \boldsymbol{0}. \tag{4.2}$$

where $\rho(\boldsymbol{r})$ is the charge density. In the previous chapter we discovered that $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$ where Φ is the electric potential. If we can determine Φ for a static distribution of charge then it is a simple matter to compute $\boldsymbol{E}$. But what are the equations satisfied by Φ?

4.1 Poisson and Laplace's Equations

Substituting $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$ into equation (4.1) we obtain

$$\boldsymbol{\nabla}^2\Phi = -\frac{\rho}{\epsilon_0}. \tag{4.3}$$

This equation is called Poisson's equation, and in regions where the charge density is zero it simplifies to Laplace's equation

$$\boldsymbol{\nabla}^2\Phi = 0. \tag{4.4}$$

Equation (4.2) is automatically satisfied for $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$ and gives us no further constraint on Φ. Consequently problems in electrostatics will come down to solving Poisson's equation (4.3) for Φ. Abstractly, this is a problem we realise that we have already solved in equation (3.6), which we reproduce here:

$$\Phi = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|} dv'.$$

This is very interesting as it demonstrates a general solution to the Poisson equation, the only problem is that the integral above may be very difficult to solve to find an explicit expression

for Φ. We will revisit this idea later in this chapter. We will need some alternative methods to determine Φ, and we will begin by studying solutions to Laplace's equation (4.4).

Laplace's equation is valid in regions where the charge density is zero, and for every electric field we have studied so far there are vast regions in which the electric field is not zero but the charge density is zero: the electric field is associated with charge located elsewhere. So in every interesting scenario there will be regions where Laplace's equation will hold.

4.1.1 Harmonic Functions

Functions which solve Laplace's equation are called harmonic functions. For example in one dimension $H_1(x) = ax + b$ is a harmonic function, in two dimensions $H_2(x, y) = \frac{1}{2}\ln(x^2 + y^2) = \ln(r)$ is a harmonic function and in three dimensions $H_3(x, y, z) = \frac{1}{\sqrt{x^2+y^2+z^2}} = \frac{1}{r}$ is an example of a harmonic function.

Comment(s).

- *If Φ satisfies $\boldsymbol{\nabla}^2\Phi = 0$ in some region V with boundary S then by Green's theorem,*

$$0 = \int_V \boldsymbol{\nabla}^2\Phi \, dv = \oint_S \boldsymbol{\nabla}\Phi \cdot d\boldsymbol{a} \tag{4.5}$$

  which is the integral form of Gauss's law for a region of zero net charge, but now written in terms of Φ.

- *Green's second identity*[1] *for two scalar functions f and g is*

$$\int_V (f\boldsymbol{\nabla}^2 g - g\boldsymbol{\nabla}^2 f)dv = \oint_S (f\boldsymbol{\nabla}g - g\boldsymbol{\nabla}f) \cdot d\boldsymbol{a}.$$

  For the case where $f = \frac{1}{r}$ and $g = \Phi$ in a region V which is a ball of radius R and where $\boldsymbol{\nabla}^2\Phi = 0$, this gives

$$-\int_V \Phi \, \boldsymbol{\nabla}^2\Big(\frac{1}{r}\Big)dv = \oint_S \Big(\frac{1}{R}\boldsymbol{\nabla}\Phi + \Phi\frac{1}{R^2}\,\boldsymbol{e_r}\Big) \cdot d\boldsymbol{a} \tag{4.6}$$

  where we have used $\boldsymbol{\nabla}(\frac{1}{r}) = -\frac{1}{r^2}\,\boldsymbol{e_r}$, which on the sphere S becomes $-\frac{1}{R^2}\,\boldsymbol{e_r}$. Now the divergence of $\frac{1}{r^2}\,\boldsymbol{e_r}$ was a function we investigated in section 2.2.5 where we found that

$$\boldsymbol{\nabla} \cdot (\frac{1}{r^2}\,\boldsymbol{e_r}) = 4\pi\delta^3(\boldsymbol{r})$$

  and hence

$$\boldsymbol{\nabla}^2\Big(\frac{1}{r}\Big) = -4\pi\delta^3(\boldsymbol{r}).$$

[1] To be proved in problem set 2.

Substituting this into equation (4.6) *we find*

$$4\pi \int_V \Phi\, \delta^3(\boldsymbol{r}) dv = \oint_S \left(\frac{1}{R} \boldsymbol{\nabla} \Phi + \Phi \frac{1}{R^2} \boldsymbol{e_r} \right) \cdot d\boldsymbol{a}$$

On the sphere $d\boldsymbol{a} = da\, \boldsymbol{e_r}$*, where* $da = R^2 \cos\phi \, d\theta \, d\phi$*, and we also note that* $\oint_S \left(\frac{1}{R} \boldsymbol{\nabla} \Phi \right) \cdot d\boldsymbol{a} = 0$ *by equation* (4.5)*, hence we obtain,*

$$\Phi(\boldsymbol{0}) = \frac{1}{4\pi R^2} \oint_S \Phi \, da.$$

Putting this into words, if Φ *satisfies Laplace's equation over a ball of radius* R *the value of* Φ *at the origin is equal to the average of* Φ *over the surface of the ball. Of course we may translate this equation to find*

$$\Phi(\boldsymbol{r}) = \frac{1}{4\pi R^2} \oint_S \Phi \, da$$

where S *is a sphere of radius* R *centred at* $\boldsymbol{r}$*. The key idea to take away from this, which is true for any harmonic function, is that* $\Phi(\boldsymbol{r})$ *is equal to the average of the values of* Φ *over a sphere around* $\boldsymbol{r}$*.*

- Φ *has no local maxima or minima over the region* V *where Laplace's equation is satisfied: any extrema of* Φ *must occur at the boundaries of this region. Suppose, to the contrary, that* Φ *had a local maximum at* $\boldsymbol{r} = \boldsymbol{r_0} \in V$ *then at all points on a sphere about* $\boldsymbol{r_0}$ *(but within* V*), the value of* Φ *would be less than* $\Phi(\boldsymbol{r_0})$*, but since* $\Phi(\boldsymbol{r_0})$ *is equal to its average values on any sphere surrounding it then this is a contradiction. Consequently* Φ *has no local maxima or minima, and any extreme values of* Φ *must occur on the boundary* S *of* V*.*

- *If* Φ *satisfies Laplace's equation in a region* V *and if, in addition,* Φ *is constant on the boundary of* V*, then* Φ *must be equal to the constant value over the entire region (otherwise there would be a local maximum or minimum within* V*).*

4.2 Boundary Conditions

Ordinary differential equations of second-order require two initial or boundary conditions before a unique solution to the equation might be found. The Laplace and Poisson equations are both second-order partial differential equations, and as such require an infinite number of boundary conditions before a unique solution to either equation might be specified[2] The customary

[2]In general one must specify both the value of the function along a boundary as well as its derivative normal to the boundary - these are called Cauchy boundary conditions, and solving a partial differential equation

boundary conditions usually given for the Laplace and Poisson equations are either the values of the function itself on a boundary (these are known as Dirichlet boundary conditions) or the values of the normal derivative of the function along a boundary (these are known as Neumann boundary conditions).[3]

To find a solution to a partial differential equation with n variables one must specify boundary conditions on an $(n-1)$ dimensional surface. For the Poisson and Laplace equation in three dimensions, the boundary conditions will need to be specified on a two dimensional surface. We are already a little familiar with cases where we have effectively solved the Poisson equation for Φ when the potential is specified on a surface, or the charge density is known (so that the potential may be computed on a surface). Furthermore if we recall that since $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$ then the value of the electric field on a boundary specifies the normal derivative of the potential, and this idea of specifying the electric field on a surface in order to compute Φ is a typical problem in electrostatics.

4.2.1 The Uniqueness of Solutions to Laplace's Equation

Theorem 4.1. *The solution to Laplace's equation in some volume V is uniquely determined if Φ is specified on the boundary S.*

Proof. We will give a proof by contradiction: suppose that there exist two solutions $\Phi_1 \neq \Phi_2$ to Laplace's equation:

$$\boldsymbol{\nabla}^2\Phi_1 = 0 \quad \text{and} \quad \boldsymbol{\nabla}^2\Phi_2 = 0$$

which both satisfy the boundary condition that the potential must equal Φ_S on the boundary S: $\Phi_1|_S = \Phi_S$ and $\Phi_2|_S = \Phi_S$. Now consider the function which is the difference between Φ_1 and Φ_2 namely $\Phi_D = \Phi_1 - \Phi_2$. This function must also satisfy Laplace's equation as

$$\boldsymbol{\nabla}^2\Phi_D = \boldsymbol{\nabla}^2\Phi_1 - \boldsymbol{\nabla}^2\Phi_2 = 0.$$

In addition its value on the boundary is zero as

$$\Phi_D|_S = \Phi_1|_S - \Phi_2|_S = \Phi_S - \Phi_S = 0.$$

Since Φ_D is a harmonic function which is constant on a boundary surface, then it must be constant and equal to zero throughout the surface (since a harmonic function only has extrema on S and not on its interior V). Therefore throughout V we have

$$\Phi_D = \Phi_1 - \Phi_2 = 0$$

equation with these boundary conditions is known as solving a Cauchy problem. The Laplace and Poisson equations happen to be require less boundary conditions in order to be solved.

[3]Other types of boundary conditions are also available.

and hence we find the contradiction that $\Phi_1 = \Phi_2$ throughout V. Hence Laplace's equation for a given boundary condition has a unique solution. □

Note that a very similar argument holds for solutions to Poisson's equation. As suppose that there were two potential functions which were equal to Φ_S on the boundary S but which both satisfied Poisson's equation:

$$\boldsymbol{\nabla}^2\Phi_1 = -\frac{\rho}{\epsilon_0} \quad \text{and} \quad \boldsymbol{\nabla}^2\Phi_2 = -\frac{\rho}{\epsilon_0}$$

for some charge distribution ρ in V. Now the difference satisfies Laplace's equation $\boldsymbol{\nabla}^2(\Phi_1 - \Phi_2) = 0$ and so $\Phi_1 - \Phi_2$ is a harmonic function which vanishes on the surface S, hence it is constant and equal to zero on V, and again $\Phi_1 = \Phi_2$. In summary, for a given boundary condition on the surface of a volume that contains charge or not, the potential Φ is unique.

We mentioned that the boundary condition could be specifying Φ on the surface S (Dirichlet boundary conditions) or that it could be the derivative of Φ normal to the surface $\boldsymbol{n} \cdot \boldsymbol{\nabla}\Phi$ where $\boldsymbol{n}$ is a unit vector normal to the surface (Neumann boundary conditions). How are these conditions related? Returning to the arguments above, to the point where we arrived at $\boldsymbol{\nabla}^2\Phi_D = \boldsymbol{\nabla}^2\Phi_1 - \boldsymbol{\nabla}^2\Phi_2 = 0$. Using Green's theorem for $\Phi_D\boldsymbol{\nabla}\Phi_D$ we have

$$\int_V \boldsymbol{\nabla} \cdot (\Phi_D\boldsymbol{\nabla}\Phi_D)dv = \oint_S \Phi_D\boldsymbol{\nabla}\Phi_D \cdot d\boldsymbol{a}$$

where $d\boldsymbol{a} = da \cdot \boldsymbol{n}$. Now since $\boldsymbol{\nabla} \cdot (\Phi_D\boldsymbol{\nabla}\phi_D) = (\boldsymbol{\nabla}\Phi_D)^2 + \Phi_D\boldsymbol{\nabla}^2\Phi_D$ and because $\boldsymbol{\nabla}^2\Phi_D = 0$ throughout V then we have

$$\int_V (\boldsymbol{\nabla}\Phi_D)^2 dv = \oint_S \Phi_D\boldsymbol{\nabla}\Phi_D \cdot \boldsymbol{n}\, da$$

Now on the surface S if the Neumann boundary conditions are specified, i.e. $\boldsymbol{n}\cdot\boldsymbol{\nabla}\Phi_1|_S = \boldsymbol{n}\cdot\boldsymbol{\nabla}\Phi|_S$ and $\boldsymbol{n} \cdot \boldsymbol{\nabla}\Phi_2|_S = \boldsymbol{n} \cdot \boldsymbol{\nabla}\Phi|_S$, as both potentials Φ_1 and Φ_2 must possess the same derivative in the direction normal to S, then $\boldsymbol{n} \cdot \boldsymbol{\nabla}\Phi_D = 0$. The right-hand-side of the equation above is zero and we find that

$$\int_V (\boldsymbol{\nabla}\Phi_D)^2 dv = 0.$$

From which we conclude that $\boldsymbol{\nabla}\Phi_D = \boldsymbol{0}$ which tells us only that the potentials Φ_1 and Φ_2 differ by at most a constant. This conclusion, reached by suing Neumann boundary conditions, is slightly weaker than but consistent with the conclusion we reached using Dirichlet boundary conditions, where we found that $\Phi_1 = \Phi_2$.

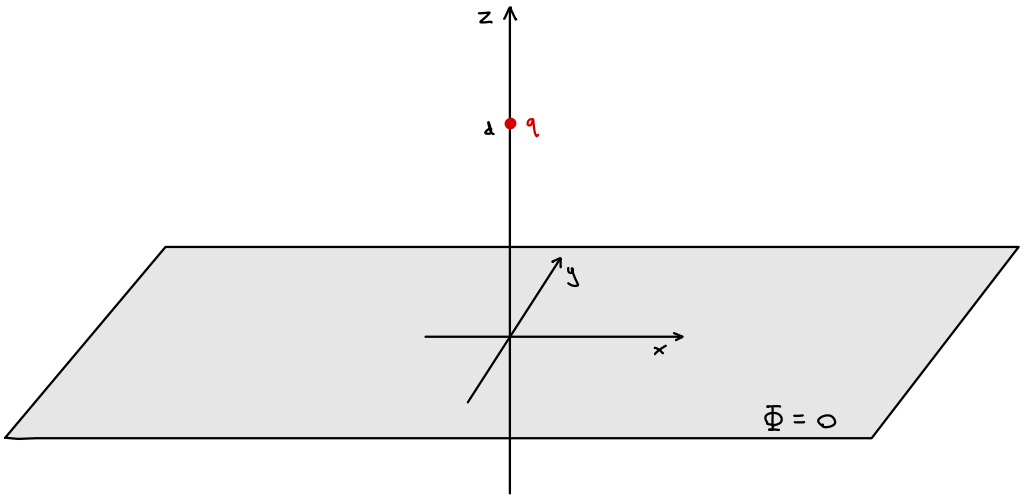
A very useful consequence of the uniqueness of solutions for given boundary conditions, is that it suffices to simply find one solution. While this is self-evident, it does mean that on occasion one will be able to guess an obvious solution to such a problem, and then will realise it is the only solution to a problem. In this setting, clever guessing of solutions is completely rewarded. Another nice consequence of the uniqueness theorems is called the method of images.

4.2.2 The Method of Images

The idea behind the method of images, is that there are situations in which we wish to find the potential Φ_V subject to some conditions on a boundary S, in which those boundary conditions could be mimicked and exactly reproduced by enlarging the setting of the problem to V' and adding some extra charges. When this is the case, especially when there is some symmetry to the array of charge overall, one can sometimes rapidly solve the enlarged problem to find the overall potential $\Phi_{V'}$. Since there is a region of the enlarged problem containing exactly the initial problem, with the same boundary conditions, then we can restrict $\Phi_{V'}$ to V to find $\Phi_V = \Phi_{V'}|_V$ because the potential is unique for a given set of boundary conditions. We will illustrate this method with a canonical example, in which the symmetry of the required construction will demonstrate why the technique is called the method of images.

Example 4.1. *Consider a point charge q located a distance d above an infinite conducting plane, whose potential is maintained at zero (this is called a grounded plane[4]). What is the potential in the region above the plane?*

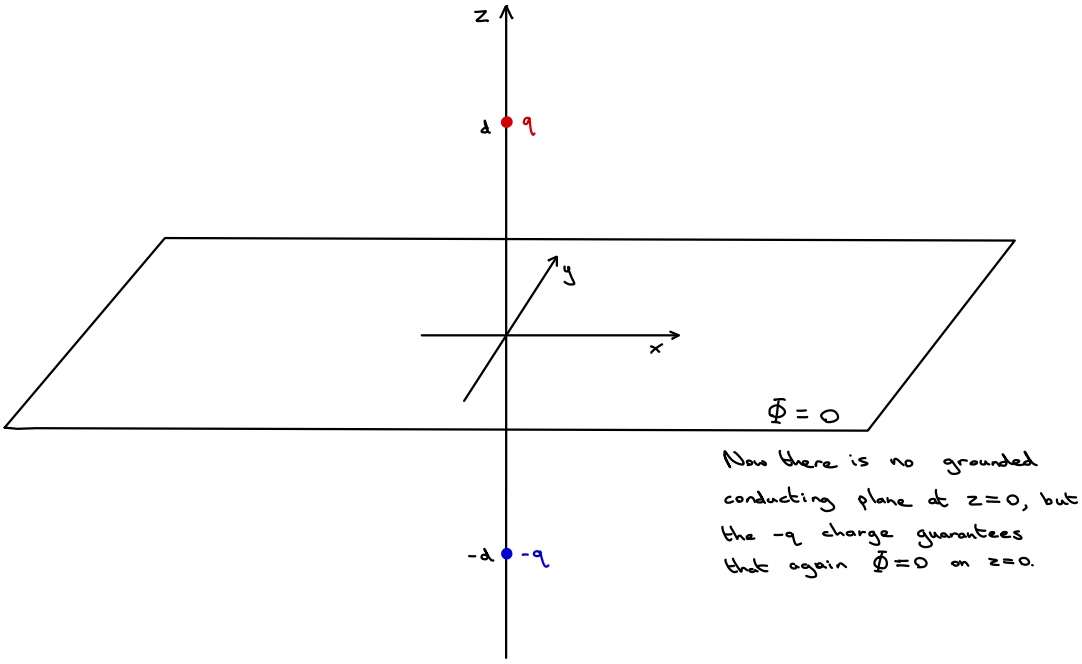
We may choose the grounded conducting plane to be the xy-plane, and let the charge q have coordinates $(0, 0, d)$. We must find the potential in the region $z > 0$.



This is an example where we know the potential Φ on the boundary of the problem, namely on the plane where it is zero and also at infinity (the remaining boundary of the region $z > 0$)

[4]The idea of mathematically setting a local region to zero potential is an interesting one: it is as if infinity could be made local. Grounding is something that can be simply achieved in real life, but mathematically it represents something very interesting. We may also say that a conductor is 'connected to ground' or just 'grounded', in all cases it means we choose the potential to be zero on the conductor.

where it is also zero. There is a point charge in this region so we are aiming to solve Poisson's equation $\boldsymbol{\nabla}^2\Phi = -\frac{\rho}{\epsilon_0}$ *where* ρ *is the charge distribution for the point charge* q *at* $(0,0,d)$*. Now we know that* $\rho = q\delta(x)\delta(y)\delta(z-d)$ *so we could set about trying to solve Poisson's equation directly. Instead we are going to change the problem in a way that will make it simple to find the potential. If, instead, we had been thinking about the problem of finding the potential when there are two charges: one of charge* q *at* $(0,0,d)$ *and a second of charge* $-q$ *at* $(0,0,-d)$ *(the mirror image of the position of* q *in the conducting plane - in the original problem) then we would have found that the potential was zero again along the xy-plane.*



How do we know that the potential is zero on $z = 0$*? It is a simple matter for us to sum the potentials for the two individual charges, to find the total potential. We first recall that the potential for a point charge* q *at* $\boldsymbol{r'}$ *is* $\Phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\frac{q}{|\boldsymbol{r}-\boldsymbol{r'}|}$ *hence the sum of the potential for charges* $\pm q$ *on the z-axis at* $z = \pm d$ *is*

$$\Phi(x,y,z) = \frac{1}{4\pi\epsilon_0}\Big(\frac{q}{\sqrt{x^2+y^2+(z-d)^2}} - \frac{q}{\sqrt{x^2+y^2+(z+d)^2}}\Big). \tag{4.7}$$

Hence $\Phi(x,y,0) = 0$*. Now we notice that* $\Phi \to 0$ *at infinity, so that the boundary conditions of this potential match exactly the boundary conditions in our initial problem set in the region* $z > 0$*. Since the potential for these boundary conditions is unique then we deduce that the potential for the point charge* q *located at* $(0,0,d)$ *above a grounded infinite conducting plane at* $z = 0$ *is simply* $\Phi(x,y,z)$ *in* (4.7) *when* $z > 0$*.*

We already know enough about the arrangement in the example above to make a claim about the charge that is induced on the grounded charge plane. First imagine that instead of an infinite conducting plane we had a sphere of radius R about the charge q, so that the charge q sits inside a cavity in a conductor. By our earlier observations about conductors we would expect a charge $-q$ to be induced on the inside of the conducting shell. Now imagine that R was allowed to grow: the curving shell would approximately seem flat, but it would still carry a charge of $-q$. This is a rough argument but we may use the potential found above to compute precisely the charge density on the plane.

We must say a little bit more about the meaning of grounding a conductor. The infinite conducting plane is grounded setting its potential to zero, and so its electric field is zero. Grounding theoretically means that an infinite supply of negative charge is able to move onto the plane. Consider a grounded spherical conducting shell surrounding a point charge q. The electric field of the charge q induces a charge $-q$ distributed over the inside of the shell, and in a neutral conductor that would result in a charge $+q$ on the outside of the shell. However when a conductor is grounded any positive charge will be cancelled out by a never ending supply of negative charge. The surface has a net charge of $-q$ and any field lines from the charge q in the interior of the shell will terminate on the shell. Outside the grounded conductor the electric field is zero. It is a similar situation with the grounded infinite plane. The electric field is non-zero above the plane but is zero below the plane.

The grounded infinite conducting plane is an equipotential, and so field lines from the point charge q are at right-angles to the plane. The discontinuity in the electric field at the grounded conducting plane is

$$\lim_{z\to 0}\Big[\boldsymbol{E}_{z>0}-\boldsymbol{E}_{z<0}\Big]=\frac{\sigma}{\epsilon_0}$$

which we computed in equation (3.5) and where σ is the charge density on the plane. Since $\boldsymbol{E}_{z<0}=\boldsymbol{0}$ then we know that at the plane's surface we have $\boldsymbol{E}|_{z\to 0^+}=\frac{\sigma}{\epsilon_0}\boldsymbol{e_z}$. Consequently, in the limit $z\to 0^+$, we have $-\boldsymbol{\nabla}\Phi=\frac{\sigma}{\epsilon_0}\boldsymbol{e_z}$, hence

$$-\frac{\partial\Phi}{\partial z}\Big|_{z\to 0^+}=\frac{\sigma}{\epsilon_0}.$$

Then from equation (4.7) we can compute $\frac{\partial\Phi}{\partial z}\Big|_{z\to 0^+}$ to obtain

$$-\frac{\partial\Phi}{\partial z}\Big|_{z\to 0^+}=-\frac{1}{2\pi\epsilon_0}\Big(\frac{qd}{(x^2+y^2+d^2)^{\frac{3}{2}}}\Big)=\frac{\sigma}{\epsilon_0}.$$

Then we may integrate $\sigma(x,y)$ obtained above over the xy-plane to find the total charge on the plane:

$$\int_{\mathbb{R}^2}\sigma\,da=-\frac{1}{2\pi}\int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty}\frac{qd}{(x^2+y^2+d^2)^{\frac{3}{2}}}\,dx\,dy$$

We may use polar coordinates in the plane to simplify the integral i.e. $x = r\cos\theta$, $y = r\sin\theta$, then we compute

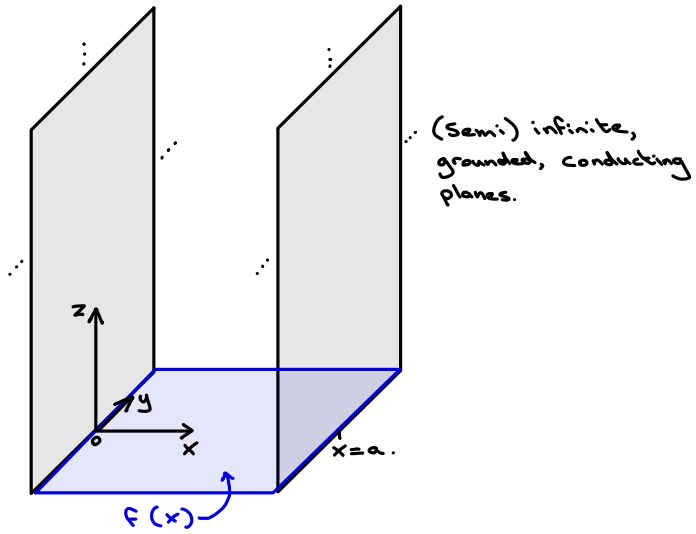
$$-\frac{1}{2\pi}\int_{r=0}^{\infty}\int_{\theta=0}^{2\pi}\frac{qd}{(r^2+d^2)^{\frac{3}{2}}}\,r\,dr\,d\theta = -\int_{r=0}^{\infty}\frac{d}{dr}\Big(\frac{-qd}{(r^2+d^2)^{\frac{1}{2}}}\Big)\,dr = -q$$

as we had anticipated.

4.3 Separation of Variables in Laplace's Equation

The circumstances in which one can make quick use of the method of images will be limited, so how are we to solve Laplace's equation and Poisson's equations in more general circumstances. We are aided by the knowledge that solutions to each equation for given boundary conditions are unique, so we may be optimistic and try out standard, simple methods to solve partial differential equations. Any solution we might find will be the unique one. It will turn out that separation of variables will be very helpful to us in solving the Laplace equation.

Example 4.2. *Consider two (semi) infinite grounded conducting planes, one coincident with the plane $x = 0$ for $z > 0$ and another coincident with $x = a$ for $z > 0$. The separation between the two planes is a and in the plane $z = 0$ between the two planes (but insulated from them) a strip (of width a) is maintained at a potential $f(x)$:*



Find the potential between the planes in the region $z > 0$.

We must solve Laplace's equation between the two planes. In the region $z > 0$ notice that due to the symmetry of the problem the electric field and the potential must be independent of

y. Hence Laplace's equation in Cartesian coordinates is

$$\frac{\partial^2\Phi}{\partial x^2}+\frac{\partial^2\Phi}{\partial z^2}=0$$

which we must solve for the boundary conditions on each side of the volume contained between the planes, for $z>0$ these are:

$$\begin{aligned}\Phi(0,z)=\Phi(a,z)&=0\\ \Phi(x,0)&=f(x)\\ \lim_{z\to\infty}[\Phi(x,z)]&=0\end{aligned}$$

The last boundary condition on Φ represents our normalisation that chooses $\Phi=0$ at infinity.

We will try to solve this problem by considering potentials of the form

$$\Phi(x,z)=X(x)Z(z).$$

Substituting this into Laplace's equation and then dividing through by $X(x)Y(y)Z(z)$ we obtain

$$\frac{1}{X}\frac{\partial^2X}{\partial x^2}+\frac{1}{Z}\frac{\partial^2Z}{\partial z^2}=0$$

where both of the terms depend on a different variable, and yet their sum is always zero. Consequently each term must be constant and we may deduce that

$$\frac{\partial^2X}{\partial x^2}=k_1X,\quad \textit{and}\quad \frac{\partial^2Z}{\partial z^2}=k_2Z,$$

where k_1 and k_2 are constants such that $k_1+k_2=0$. The solutions to these equations will be trigonometric or hyperbolic functions, depending upon the signs of k_1 and k_2. The boundary conditions for Φ give boundary conditions for X and Z. Substitution of $\Phi=XZ$ into the boundary conditions yields

$$\begin{aligned}X(0)=X(a)&=0,\\ X(x)Z(0)&=f(x),\quad \textit{and}\\ \lim_{z\to\infty}[Z(z)]&=0.\end{aligned}$$

The second condition implies that $Z(0)$ is a non-zero constant, which we may choose to normalise to one, so that

$$Z(0)=1\quad \textit{and}\quad X(x)=f(x).$$

Since $X(0)=0=X(a)$ then X must be a sinusoidal function with a zero at $x=a$, i.e. we find that, up to constant coefficients there are multiple solutions labelled by $n\in\mathbb{Z}$:

$$X_n(x)=\sin(\frac{n\pi}{a}x)\quad\Longrightarrow\quad k_1=-\Big(\frac{n\pi}{a}\Big)^2=:-k_n.$$

Since $k_2 = -k_1 = k_n = \left(\frac{n\pi}{a}\right)^2$ *is positive, then we surmise that* Z *is a combination of hyperbolic functions satisfying* $Z(0) = 1$ *such that* $\lim_{z\to\infty}(Z(z)) = 0$. *Writing*

$$Z_n(z) = B\sinh(\sqrt{k_n}z) + C\cosh(\sqrt{k_n}z)$$

where we have replaced Z *with* Z_n *to indicate that there are multiple solutions each one labelled by* n. *The boundary conditions mean that the function is neither hyperbolic sine or hyperbolic cosine, hence we rewrite our function in terms of its constituent exponential functions*

$$Z_n(z) = \frac{1}{2}(B+C)e^{\sqrt{k_n}z} + \frac{1}{2}(-B+C)e^{-\sqrt{k_n}z}$$

Then the boundary condition $\lim_{z\to\infty}[Z_n(z)] = 0$ *implies that* $B = -C$ *and the boundary condition that* $Z_n(0) = 1$ *leads us to* $-B + C = 2C = 2$ *hence*

$$Z_n(z) = e^{-\sqrt{k_n}z} = e^{-\frac{n\pi}{a}z}$$

Now any linear sum of the functions $X_n(x)Z_n(z)$ *will solve Laplace's equation so we have deduced that*

$$\Phi(x,z) = \sum_{n=1}^{\infty} C_n e^{-\frac{n\pi}{a}z}\sin(\frac{n\pi}{a}x)$$

solves Laplace's equation for any choice of constant coefficients C_n. *We have not yet imposed the boundary condition that* $\Phi(x,0) = f(x)$ *which gives us*

$$\Phi(x,0) = \sum_{n=1}^{\infty} C_n\sin(\frac{n\pi}{a}x) = f(x).$$

We may invert this expression to find unique expressions for C_n *for any given* $f(x)$. *The ability to do this rests upon the observation that*

$$I_{nm} = \int_0^{\pi}\sin(nx)\sin(mx)dx = \delta_{nm}\frac{\pi}{2}$$

which is proved in appendix D. We may make a substitution to find a form of this expression appropriate for a different periodic length, for the case in hand in the x*-direction has boundaries at* 0 *and* a, *rather than* 0 *and* π, *and by substituting* $x' = \frac{a}{\pi}x$ *we obtain*

$$\int_0^{a}\sin(\frac{n\pi}{a}x')\sin(\frac{m\pi}{a}x')dx' = \delta_{nm}\frac{a}{2}.$$

Now by considering the integral

$$\begin{aligned}\int_0^a f(x)\sin(\frac{m\pi}{a}x)dx &= \int_0^a\sum_{n=1}^{\infty}C_n\sin(\frac{n\pi}{a}x)\sin(\frac{m\pi}{a}x)\,dx\\ &= \sum_{n=1}^{\infty}C_n\delta_{nm}\frac{a}{2}\\ &= \frac{a}{2}C_m\end{aligned}$$

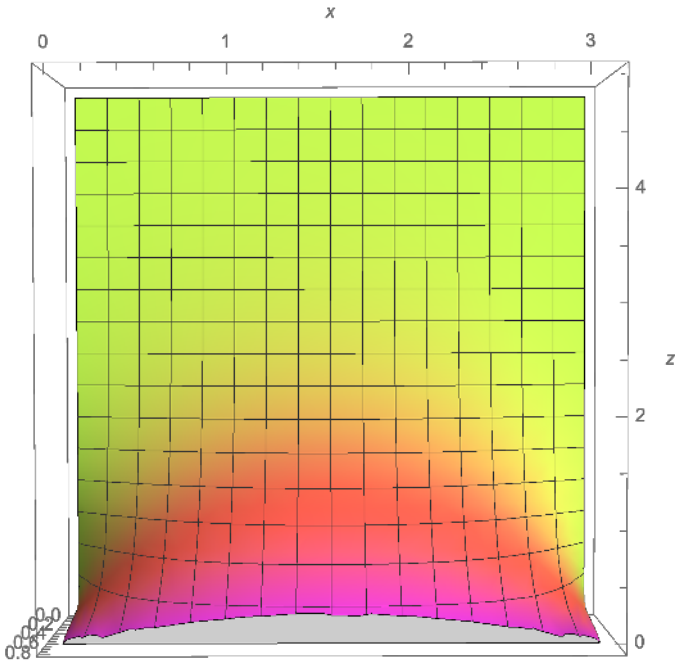
we find a unique expression for C_n in terms of $f(x)$. Once we are given an explicit expression for $f(x)$ we may find Φ explicitly. For example suppose that $f(x) = K$, a constant. Then

$$\begin{aligned} C_m &= \frac{2K}{a}\int_0^a \sin(\frac{m\pi}{a}x)dx \\ &= \frac{2K}{a}\int_0^a \frac{d}{dx}\Big(-\frac{a}{m\pi}\cos(\frac{m\pi}{a}x)\Big)dx \\ &= \begin{cases} 0 & \text{even } m \\ \frac{4K}{m\pi} & \text{odd } m. \end{cases} \end{aligned}$$

Therefore for $f(x) = K$ we find

$$\Phi(x,z) = \sum_{n=1,3,5,\ldots}^{\infty} \frac{4K}{n\pi} e^{-\frac{n\pi}{a}z}\sin(\frac{n\pi}{a}x).$$

We plot the first ten terms of the function for $a = \pi$ and $K = 1$ below



at $z = 0$ the potential is equal to K (shown here for $K = 1$) and is zero on the planes at $x = 0$ and $x = a$ (shown here for $a = \pi$).

In the example above we have shown how Laplace's equation may be solved in Cartesian coordinates, but frequently we may wish to work in spherical coordinates where the orthogonal functions needed to solve the problem will be different.

There are similar arguments that can be made in spherical coordinates, and the orthogonal polynomials which appear are related to the Legendre polynomials. The discussion of the solutions to the Laplace equations becomes much more involved in these coordinates and many interesting ideas turn up - the interested reader may read a little about the development of the classical orthogonal polynomials in appendix D, however a full discussion of these ideas will require a journey that takes one far from the discussion of electromagnetism. Consequently we will simply review the key ideas here, and state some of the results, which may prove useful.

In spherical polar coordinates, we customarily use the angular coordinate ϕ, so in this setting we will denote the potential by Ψ, and its gradient is:

$$\boldsymbol{\nabla}\Psi = \frac{\partial\Psi}{\partial r}\,\boldsymbol{e_r} + \frac{1}{r\cos\phi}\frac{\partial\Psi}{\partial\theta}\,\boldsymbol{e_\theta} + \frac{1}{r}\frac{\partial\psi}{\partial\phi}\,\boldsymbol{e_\phi}.$$

To find the Laplacian we first compute

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\boldsymbol{e_r} = \cos\phi\,\boldsymbol{e_\theta} && \frac{\partial}{\partial\theta}\boldsymbol{e_\theta} = \sin\phi\,\boldsymbol{e_\phi} - \cos\phi\,\boldsymbol{e_r} && \frac{\partial}{\partial\theta}\boldsymbol{e_\phi} = -\sin\phi\,\boldsymbol{e_\theta}\\
&\frac{\partial}{\partial\phi}\boldsymbol{e_r} = \boldsymbol{e_\phi} && && \frac{\partial}{\partial\phi}\boldsymbol{e_\phi} = -\boldsymbol{e_r}
\end{aligned}$$

with all other derivatives of $\boldsymbol{e_r}$, $\boldsymbol{e_\theta}$ and $\boldsymbol{e_\phi}$ being zero. Consequently, we compute[5] that Laplace's equation for Ψ is

$$\boldsymbol{\nabla}^2\Psi = \frac{1}{r^2}\frac{\partial}{\partial r}\Big(r^2\frac{\partial\Psi}{\partial r}\Big) + \frac{1}{r^2\cos^2\phi}\frac{\partial^2\Psi}{\partial\theta^2} + \frac{1}{r^2\cos\phi}\frac{\partial}{\partial\phi}\Big(\cos\phi\frac{\partial\Psi}{\partial\phi}\Big) = 0.$$

To separate the variables we split the potential into a product of functions as

$$\Psi(r,\theta,\phi) = R(r)\Theta(\theta)\Phi(\phi)$$

then substitution into Laplace's equation followed by multiplication by $\frac{r^2}{\Psi}$ gives:

$$\frac{1}{R}\frac{\partial}{\partial r}\Big(r^2\frac{\partial R}{\partial r}\Big) + \frac{1}{\cos^2\phi}\frac{1}{\Theta}\frac{\partial^2\Theta}{\partial\theta^2} + \frac{1}{\cos\phi}\frac{1}{\Phi}\frac{\partial}{\partial\phi}\Big(\cos\phi\frac{\partial\Phi}{\partial\phi}\Big) = 0. \tag{4.8}$$

Since the equation above is complicated we will first consider a simplified version: we will consider potentials which are independent of θ, i.e. we will set $\Theta(\theta)$ to be constant for this first discussion of Laplace's equation in spherical coordinates. In this case we then have

$$\frac{1}{R}\frac{\partial}{\partial r}\Big(r^2\frac{\partial R}{\partial r}\Big) + \frac{1}{\cos\phi}\frac{1}{\Phi}\frac{\partial}{\partial\phi}\Big(\cos\phi\frac{\partial\Phi}{\partial\phi}\Big) = 0.$$

[5]Remember to allow any derivative to act on a vector before taking the dot product when calculating $\boldsymbol{\nabla}\cdot\boldsymbol{\nabla} = \boldsymbol{\nabla}^2$, before collecting together the six non-zero terms into the form shown here.

Our simplification has led to the two terms depending on different coordinates, and consequently they must each be separately constant. We will choose the constant to be an even number such that

$$\frac{1}{R}\frac{\partial}{\partial r}\Big(r^2\frac{\partial R}{\partial r}\Big) = l(l+1)$$
$$\frac{1}{\cos\phi}\frac{1}{\Phi}\frac{\partial}{\partial\phi}\Big(\cos\phi\frac{\partial\Phi}{\partial\phi}\Big) = -l(l+1)$$

for $l \in \mathbb{Z}^+$. We have chosen the constants to have this form because the second equation has solutions which are given by the Legendre polynomials with $\Phi(\phi) = P_l(\sin\phi)$ where

$$P_l(x) = \frac{1}{2^l l!}\frac{d^l}{dx^l}(x^2-1)^l.$$

One can read a brief summary of interesting statements about the Legendre polynomials in appendix D - and in particular note that they form a complete, orthogonal basis for the polynomial functions just as the functions $\sin(nx)$ do, properties that proved essential when solving the Laplace equation in Cartesian coordinates.

We may solve the equation for $R(r)$ to find

$$R(r) = Ar^l + \frac{B}{r^{l+1}}$$

where A and B are constants to be fixed by the boundary conditions of the problem. Consequently we find the infinite set of solutions (which are independent of θ) are

$$\Psi(r,\phi) = \sum_{l=0}^{\infty}\Big(A_l r^l + \frac{B_l}{r^{l+1}}\Big)P_l(\sin\phi).$$

The orthogonality condition for the Legendre polynomials when the argument is $\sin\phi$ rather than x becomes

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} P_n(\sin\phi)P_m(\sin\phi)\cos\phi\,d\phi = \frac{2}{2n+1}\delta_{nm}$$

i.e. one must integrate against a measure $\cos\phi\,d\phi$ to take advantage of the orthogonality relation, and for a given set of boundary conditions then determine the unique solution by finding A_l and B_l.

The solution to the Laplace equation when Φ depends on θ is similar. Now we contemplate equation (4.8) again, but commence with the term we eliminated above $\frac{1}{\cos^2\phi}\frac{1}{\Theta}\frac{\partial^2\Theta}{\partial\theta^2}$. By choosing the θ parts of this term to be equal to a constant such that

$$\frac{1}{\cos^2\phi}\frac{1}{\Theta}\frac{\partial^2\Theta}{\partial\theta^2} = -\frac{m^2}{\cos^2\phi}$$

for $m \in \mathbb{Z}$ we extract that

$$\Theta_m(\theta) = e^{im\theta}.$$

Substituting this into (4.8) we obtain

$$\frac{1}{R}\frac{\partial}{\partial r}\Big(r^2\frac{\partial R}{\partial r}\Big) = \frac{m^2}{\cos^2\phi} - \frac{1}{\cos\phi}\frac{1}{\Phi}\frac{\partial}{\partial\phi}\Big(\cos\phi\frac{\partial\Phi}{\partial\phi}\Big) = l(l+1)$$

where $l \in \mathbb{Z}$. Consequently,

$$\frac{1}{\cos\phi}\frac{1}{\Phi}\frac{\partial}{\partial\phi}\Big(\cos\phi\frac{\partial\Phi}{\partial\phi}\Big) = \frac{m^2}{\cos^2\phi} - l(l+1)$$

which can be solved for $|m| \leq l$. With these conditions on the constants this equation has solutions (for $\Phi(\phi)$) which are called the associated Legendre polynomials and denoted $P_l^m(\sin\phi)$ where

$$P_l^m(x) = (-1)^m(1-x^2)^{\frac{m}{2}}\frac{d^m}{dx^m}(P_l(x))$$

where $P_l(x)$ are the Legendre polynomials. The associated Legendre polynomials satisfy the orthogonality condition

$$\int_{-1}^{1} P_l^n P_l^m \frac{dx}{1-x^2} = \delta^{nm}\frac{(l+m)!}{m(l-m)!}$$

for positive, non-zero m and n in $\mathbb{Z}$, while if $n = m = 0$ the integral is infinite. Now the full set of possible solutions is

$$\Psi(r,\theta,\phi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\Big(A_l^m r^l + \frac{B_l^m}{r^{l+1}}\Big)P_l^m(\sin\phi)e^{im\theta}.$$

A further refinement of this solution is to present it in terms of the spherical harmonic functions.

The essential point to take away from the above discussion of the Laplace equation in spherical coordinates is that, as with the equation in Cartesian coordinates, for given boundary conditions a unique solution exists for the electric potential which may be found by separating the variables in the problem and using the properties of orthogonal polynomials to impose specific boundary conditions.

4.4 Poisson's Equation and Green's Functions

We have solved Laplace's equation to find the potential in the absence of charge, and, as we mentioned briefly at the start of this chapter, we have also been able to solve Poisson's equation. In particular we have observed that

$$\Phi = \frac{1}{4\pi\epsilon_0}\int\frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}dv'.$$

solves the equation

$$\boldsymbol{\nabla}^2\Phi = -\frac{\rho}{\epsilon_0}.$$

Let us substitute our integral expression for Φ into Poisson's equation to learn a bit more about this solution. We have

$$\begin{aligned}\boldsymbol{\nabla}^2\Phi &= \boldsymbol{\nabla}^2\Big(\frac{1}{4\pi\epsilon_0}\int\frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}dv'\Big)\\ &= \frac{1}{4\pi\epsilon_0}\int\boldsymbol{\nabla}^2\Big(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\Big)\rho(\boldsymbol{r'})dv'\\ &= -\frac{\rho}{\epsilon_0}\end{aligned}$$

where we recall that the potential is a function of $\boldsymbol{r}$, while the integral is over $\boldsymbol{r'}$ and $\boldsymbol{\nabla}^2$ acts on $\boldsymbol{r}$ and not $\boldsymbol{r'}$. By studying the terms we realise that

$$\boldsymbol{\nabla}^2\Big(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\Big) = -4\pi\delta^3(\boldsymbol{r}-\boldsymbol{r'})$$

as then

$$\boldsymbol{\nabla}^2\Phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int\boldsymbol{\nabla}^2\Big(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\Big)\rho(\boldsymbol{r'})dv' = -\frac{1}{4\pi\epsilon_0}\int 4\pi\delta^3(\boldsymbol{r}-\boldsymbol{r'})\rho(\boldsymbol{r'})dv' = -\frac{\rho(\boldsymbol{r})}{\epsilon_0}.$$

In this procedure we began by knowing the solution to $\boldsymbol{\nabla}^2\Phi = -\frac{\rho}{\epsilon_0}$ and deduced that $\boldsymbol{\nabla}^2\Big(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\Big) = -4\pi\delta^3(\boldsymbol{r}-\boldsymbol{r'})$, consistent with our observations when we first discussed the Dirac delta function. However, in other Poisson-like problems, we are likely to work in the other direction, that is given an equation of the form

$$Df(\boldsymbol{r}) = \rho(\boldsymbol{r})$$

where D stands for an abstract differential operator, we can solve for $f(\boldsymbol{r})$ if we can identify a function $G(\boldsymbol{r}-\boldsymbol{r'})$, called the Green's function for D which satisfies

$$DG(\boldsymbol{r}-\boldsymbol{r'}) = \delta^3(\boldsymbol{r}-\boldsymbol{r'}).$$

If we can identify the Green's function for the operator D then we can write down the solution for the equation as

$$f(\boldsymbol{r}) = \int G(\boldsymbol{r}-\boldsymbol{r'})\rho(\boldsymbol{r'})dv'.$$

This is a solution to the equation as

$$\begin{aligned}Df(\boldsymbol{r}) &= D\Big(\int G(\boldsymbol{r}-\boldsymbol{r'})\rho(\boldsymbol{r'})dv'\Big)\\ &= \int(DG(\boldsymbol{r}-\boldsymbol{r'}))\rho(\boldsymbol{r'})dv'\\ &= \int\delta^3(\boldsymbol{r}-\boldsymbol{r'})\rho(\boldsymbol{r'})dv'\\ &= \rho(\boldsymbol{r}).\end{aligned}$$

The idea we have outlined shifts the challenge of directly solving a differential equation to one of identifying the Green's function for an operator. In the case of Poisson's equation $D = \boldsymbol{\nabla}^2$, the Laplacian, and we can say that the Green's function for the Laplacian is $-\frac{1}{4\pi}\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}$. We may use this observation to find the potential for a given charge distribution. Consider the point charge q located at the origin, its charge density $\rho(\boldsymbol{r}) = q\delta^3(\boldsymbol{r})$ and hence

$$\begin{aligned}\Phi(\boldsymbol{r}) &= \frac{1}{4\pi\epsilon_0}\int_V \frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}dv' \\ &= \frac{1}{4\pi\epsilon_0}\int_V \frac{q\delta^3(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}dv' \\ &= \frac{1}{4\pi\epsilon_0}\frac{q}{r}\end{aligned}$$

where in carrying out the integral, the Dirac delta function has set all values of $\boldsymbol{r'}$ equal to $\mathbf{0}$, hence the denominator in the integral goes from $|\boldsymbol{r}-\boldsymbol{r'}|$ to $|\boldsymbol{r}| = r$ and recalling that the integral is over r', θ' and ϕ', emphasises that r is a constant.

Example 4.3. *Find the potential at any point on the z-axis due to a ring of charge of radius R and constant charge density λ centred about the z-axis and located in the plane $z = z_0$.*

We first find an expression for the charge density in $\mathbb{R}^3$ which we will denote $\rho(\boldsymbol{r})$.Due to the symmetry of the problem we will work in cylindrical coordinates, but due to the notation ρ being used for the charge density we will denote the radial cylindrical coordinate by r here. We may use Dirac delta functions to describe the ring of charge as

$$\rho(\boldsymbol{r}) = \lambda\delta(r-R)\delta(z-z_0).$$

Hence on the z-axis,

$$\begin{aligned}\Phi((0,0,z)) &= \frac{1}{4\pi\epsilon_0}\int_V \frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}dv' \\ &= \frac{1}{4\pi\epsilon_0}\int_{r'=0}^{\infty}\int_{\theta'=0}^{2\pi}\int_{z'=-\infty}^{\infty}\frac{\lambda\delta(r'-R)\delta(z'-z_0)}{\sqrt{r'^2+(z-z')^2}}r'\,dr'\,d\theta'\,dz' \\ &= \frac{1}{4\pi\epsilon_0}\lambda\int_{\theta'=0}^{2\pi}\int_{z'=-\infty}^{\infty}\frac{\delta(z'-z_0)}{\sqrt{R^2+(z-z')^2}}R\,d\theta'\,dz' \\ &= \frac{1}{4\pi\epsilon_0}\lambda\int_{\theta'=0}^{2\pi}\frac{R}{\sqrt{R^2+(z-z_0)^2}}\,d\theta' \\ &= \frac{\lambda}{2\epsilon_0}\frac{R}{\sqrt{R^2+(z-z_0)^2}}.\end{aligned}$$

where we have used the observation that $|\boldsymbol{r}-\boldsymbol{r'}| = \sqrt{(0-x')^2+(0-y')^2+(z-z')^2} = \sqrt{r'^2+(z-z')^2}$ *for* $\boldsymbol{r} = \begin{pmatrix} 0 \\ 0 \\ z \end{pmatrix}$. *Notice that for* $z = z_0$ *the potential at the centre of the ring is a constant, which is to be expected as the symmetry of the problem tells us that the electric field must vanish at the centre of the ring. The electric field has only a* z*-component and using* $\boldsymbol{E}(\boldsymbol{r}) = -\boldsymbol{\nabla}\Phi$ *we find*

$$\begin{aligned}\boldsymbol{E}(\boldsymbol{r}) &= -\frac{\lambda}{2\epsilon_0}\frac{\partial}{\partial z}\Big(\frac{R}{\sqrt{R^2+(z-z_0)^2}}\Big)\,\boldsymbol{e_z} \\ &= \frac{\lambda}{2\epsilon_0}\Big(\frac{R(z-z_0)}{(R^2+(z-z_0)^2)^{\frac{3}{2}}}\Big)\,\boldsymbol{e_z}\end{aligned}$$

Example 4.4. *Find the potential due to a wire on the* z*-axis running from* $z = -\frac{d}{2}$ *to* $z = \frac{d}{2}$ *carrying constant charge density* λ*.*

The charge density is

$$\rho(\boldsymbol{r}) = \lambda\delta(x)\delta(y)\big(H(z+\frac{d}{2}) - H(z-\frac{d}{2})\big)$$

where $H(x)$ *is the Heaviside step function which is defined by*

$$H(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

hence

$$H(z+\frac{d}{2}) = \begin{cases} 0 & z < -\frac{d}{2} \\ 1 & z \geq -\frac{d}{2} \end{cases} \quad \textit{and} \quad H(z-\frac{d}{2}) = \begin{cases} 0 & z < \frac{d}{2} \\ 1 & z \geq \frac{d}{2} \end{cases}$$

so that

$$H(z+\frac{d}{2}) - H(z-\frac{d}{2}) = \begin{cases} 0 & z < -\frac{d}{2} \\ 1 & -\frac{d}{2} \leq z < \frac{d}{2} \\ 0 & z \geq \frac{d}{2}. \end{cases}$$

Hence we have,

$$\begin{aligned}
\Phi((x,y,z)) &= \frac{1}{4\pi\epsilon_0}\int_V \frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}dv' \\
&= \frac{1}{4\pi\epsilon_0}\int_{x'=-\infty}^{\infty}\int_{y'=-\infty}^{\infty}\int_{z'=-\infty}^{\infty} \frac{\lambda\delta(x')\delta(y')\big(H(z'+\frac{d}{2})-H(z'-\frac{d}{2})\big)}{\sqrt{(x-x')^2+(y-y')^2+(z-z')^2}}\,dx'\,dy'\,dz' \\
&= \frac{1}{4\pi\epsilon_0}\lambda\int_{y'=-\infty}^{\infty}\int_{z'=-\infty}^{\infty} \frac{\delta(y')\big(H(z'+\frac{d}{2})-H(z'-\frac{d}{2})\big)}{\sqrt{x^2+(y-y')^2+(z-z')^2}}\,dy'\,dz' \\
&= \frac{1}{4\pi\epsilon_0}\lambda\int_{z'=-\infty}^{\infty} \frac{H(z'+\frac{d}{2})-H(z'-\frac{d}{2})}{\sqrt{x^2+y^2+(z-z')^2}}\,dz' \\
&= \frac{1}{4\pi\epsilon_0}\lambda\int_{z'=-\frac{d}{2}}^{\frac{d}{2}} \frac{1}{\sqrt{x^2+y^2+(z-z')^2}}\,dz' \\
&= -\frac{1}{4\pi\epsilon_0}\lambda\Big(\operatorname{arcsinh}\big(\frac{z-\frac{d}{2}}{\sqrt{x^2+y^2}}\big)-\operatorname{arcsinh}\big(\frac{z+\frac{d}{2}}{\sqrt{x^2+y^2}}\big)\Big) \\
&= -\frac{1}{4\pi\epsilon_0}\lambda\ln\Big(\frac{(z-\frac{d}{2})(1+\sqrt{1+\frac{x^2+y^2}{(z-\frac{d}{2})^2}})}{(z+\frac{d}{2})(1+\sqrt{1+\frac{x^2+y^2}{(z+\frac{d}{2})^2}})}\Big)
\end{aligned}$$

The integral in the fifth line can be solved by substituting $(z-z') = \sqrt{x^2+y^2}\sinh\alpha$ *for example.*

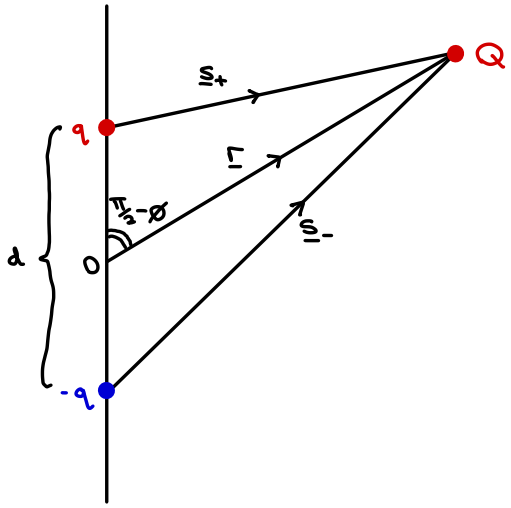
4.5 The Multipole Expansion

The electric field for point charge q at $\mathbf{0}$ has a potential at $\boldsymbol{r}$ equal to $\frac{1}{4\pi\epsilon_0}\frac{q}{r}$. If we consider a collection of point charges q_i all located at or very near the origin, then we expect the potential far from the origin to be approximately $\frac{1}{4\pi\epsilon_0}\frac{Q}{r}$ where $Q=\sum_i q_i$. But what would happen to our approximation if the sum of the charges q_i was zero? We would need another approximation, this next-to-leading order approximation will be the dipole approximation so-called as the simplest non-trivial zero charge configuration is called a dipole.

Example 4.5. *An electric dipole consists of two point charges q and $-q$ separated by a distance d. Find the approximate potential valid far from the dipole.*

Let us orient the charges so that they lie on the z-axis with the charge q at $\boldsymbol{r'_+} = \begin{pmatrix}0\\0\\\frac{d}{2}\end{pmatrix}$ *and*

the charge $-q$ *at* $\boldsymbol{r}'_- = \begin{pmatrix} 0 \\ 0 \\ -\frac{d}{2} \end{pmatrix}$, *and we will let* $\boldsymbol{s}_+ = \boldsymbol{r} - \boldsymbol{r}'_+$ *and* $\boldsymbol{s}_- = \boldsymbol{r} - \boldsymbol{r}'_-$:



To find the precise potential at $\boldsymbol{r}$ *we may sum the potentials for the two point charges to obtain*

$$\Phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\Big(\frac{q}{s_+} - \frac{q}{s_-}\Big)$$

where, using the cosine rule, we have

$$s_+^2 = r^2 + \frac{d^2}{4} - rd\cos(\frac{\pi}{2} - \phi) = r^2 + \frac{d^2}{4} - rd\sin\phi$$

and

$$s_-^2 = r^2 + \frac{d^2}{4} - rd\cos(\phi + \frac{\pi}{2}) = r^2 + \frac{d^2}{4} + rd\sin\phi$$

Altogether,

$$s_\pm^2 = r^2\Big(1 + \frac{d^2}{4r^2} \mp \frac{d}{r}\sin\phi\Big).$$

In the region where $r \gg d$ *we will treat the term* $\frac{d^2}{4r^2}$ *as negligible to make the approximation*

$$s_\pm^2 \approx r^2\Big(1 \mp \frac{d}{r}\sin\phi\Big)$$

and

$$\frac{1}{s_\pm} \approx \frac{1}{r}\Big(1 \mp \frac{d}{r}\sin\phi\Big)^{-\frac{1}{2}} \approx \frac{1}{r}\Big(1 \pm \frac{d}{2r}\sin\phi\Big)$$

where in the final approximation we have used the Taylor expansion of $f(x) = x^{-\frac{1}{2}}$ *about* $x = 1$, *i.e.* $f(1+h) = f(1) + hf'(1) + \mathcal{O}(h^2) = 1 - \frac{h}{2} + \mathcal{O}(h^2)$ *and then neglected the terms of order* $\frac{d^2}{r^2}$ *and smaller, to make the approximation. Under our approximation we have*

$$\Phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\Big(\frac{q}{s_+} - \frac{q}{s_-}\Big) \approx \frac{1}{4\pi\epsilon_0}\frac{qd\sin\phi}{r^2}.$$

The potential for the dipole, at distances $r \gg d$, goes as $\frac{1}{r^2}$, while the potential for the point charge[6] varies as $\frac{1}{r}$. If we combine a pair of equal but opposite dipoles to form a square[7] one forms a quadrupole whose potential (again for $r \gg d$) goes as $\frac{1}{r^3}$. One can continue and build arrangements of 2^n charges whose total charge is zero but whose potential varies as $\frac{1}{r^{n+1}}$ for $r \gg d$. On the other hand if we could expand a potential $\Phi(\boldsymbol{r})$ in terms of powers of $\frac{1}{r}$, then $\frac{1}{r}$ terms in the potential may be interpreted as coming from the monopole, $\frac{1}{r^2}$ the dipole approximation, $\frac{1}{r^3}$ from the quadrupole expansion and so on. Such an expansion is called the multipole expansion, we will now show that it is possible to expand a potential in this way for a general charge distribution $\rho(\boldsymbol{r})$.

Consider the potential sourced by ρ which is given by the volume integral

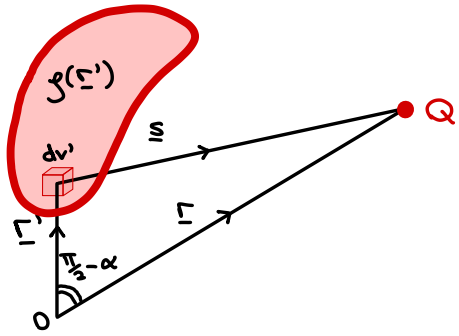
$$\Phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r'})}{s}dv'$$

where $\boldsymbol{s} = \boldsymbol{r} - \boldsymbol{r'}$. For each infinitesimal volume of charge $\rho(\boldsymbol{r'})dv'$ located at $\boldsymbol{r'}$ we may compute s for a test charge at $\boldsymbol{r}$ following a similar computation to that used to find the approximate potential for the dipole. There is a small difference to be considered: for the dipole we were able to orient the axis of the dipole along the z-axis, and then the angle between the line from the origin to either charge on the z-axis and $\boldsymbol{r}$ was directly related to our spherical coordinate ϕ. For a volume of charge we will need to consider the angle between $\boldsymbol{r'}$ and $\boldsymbol{r}$ which is not directly related to the spherical coordinate ϕ, but to make a comparison with ϕ in the dipole

[6] In this naming scheme the point charge is called the electric monopole.

[7] whose charges, at the vertices of the square, have alternating sign as one travels around the corners of the square

potential we will call this angle $\frac{\pi}{2} - \alpha$ as shown below.



The computation of s is very similar to the computation of S_+ for the dipole and we use the cosine rule to find

$$s^2 = r^2\Big(1 + \frac{r'^2}{r^2} - 2\frac{r'}{r}\sin\alpha\Big).$$

We no longer make any kind of approximation. We need to find $\frac{1}{s}$ in order to compute $\Phi(\boldsymbol{r})$, and by rearranging and then Taylor expanding[8] to obtain

$$\begin{aligned}\frac{1}{s} &= \frac{1}{r}\Big[1 - \frac{1}{2}\Big(\frac{r'^2}{r^2} - 2\frac{r'}{r}\sin\alpha\Big) + \frac{3}{8}\Big(\frac{r'^2}{r^2} - 2\frac{r'}{r}\sin\alpha\Big)^2 - \frac{5}{16}\Big(\frac{r'^2}{r^2} - 2\frac{r'}{r}\sin\alpha\Big)^3 \\ &\qquad + \frac{35}{128}\Big(\frac{r'^2}{r^2} - 2\frac{r'}{r}\sin\alpha\Big)^4 + \ldots\Big] \\ &= \frac{1}{r}\Big[1 + \frac{r'}{r}\sin\alpha + \Big(\frac{r'}{r}\Big)^2\Big(\frac{3\sin^2\alpha - 1}{2}\Big) + \Big(\frac{r'}{r}\Big)^3\Big(\frac{-3\sin\alpha + 5\sin^3\alpha}{2}\Big) \\ &\qquad + \Big(\frac{r'}{r}\Big)^4\Big(\frac{3 - 30\sin^2\alpha + 35\sin^4\alpha}{8}\Big) + \ldots\Big] \end{aligned} \tag{4.9}$$

where, in the final step, we have collected together terms as powers of $\frac{r'}{r}$ which was our aim in convincing ourselves that a multipole expansion of a general charge distribution is possible. So far we have just determined $\frac{1}{s}$, we should put this into the integral for Φ to see that we have achieved our aim.

First we note a curiosity. By studying the terms in the expansion of $\frac{1}{s}$ and comparing with the Legendre polynomials in table D.2 we realise that these first few terms contain Legendre polynomials of the form $P_n(\sin\alpha)$. This pattern continues[9] and in fact,

$$\frac{1}{s} = \frac{1}{r}\sum_{n=0}^{\infty}\Big(\frac{r'}{r}\Big)^n P_n(\sin\alpha)$$

[8] The Taylor expansion we use is the same as for the case of the dipole, but now we keep track of more terms. Explicitly we are Taylor expanding $f(x) = x^{-\frac{1}{2}}$ about $x = 1$, i.e. $f(1+h) = \sum_{n=0}^{\infty}\frac{h^n}{n!}f^n(1) = 1 - \frac{1}{2}h + \frac{3}{8}h^2 - \frac{5}{16}h^3 + \frac{35}{128}h^4 + \ldots$.

[9] $\frac{1}{s}$ is a generating function for the Legendre polynomials.

and so we obtain

$$\begin{aligned}\Phi(\boldsymbol{r}) &= \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r'})}{r}\sum_{n=0}^{\infty}\Big(\frac{r'}{r}\Big)^n P_n(\sin\alpha)dv' \\ &= \frac{1}{4\pi\epsilon_0}\Big[\int \frac{\rho(\boldsymbol{r'})}{r}dv' + \int \frac{\rho(\boldsymbol{r'})r'}{r^2}(\sin\alpha)dv' + \int \frac{\rho(\boldsymbol{r'})r'^2}{r^3}\Big(\frac{3\sin^2\alpha-1}{2}\Big)dv' + \dots\Big]\end{aligned}$$

This is the multipole expansion of the potential for a charge density $\rho(\boldsymbol{r'})$. The first term is a monopole contribution, the second a dipole contribution, the third a quadrupole contributions and so on. This means that an exact electric potential for a distribution of charge can be approximated for $r \gg r'$ by a system of multipoles constructed out of discrete point charges. Furthermore it means that normally the potential is dominated by the monopole term, but for cases where the total charge vanishes then the leading term is the dipole term. Of course the dipole term may vanish (we might say when the total dipole charge is zero) and then the leading term is the quadrupole term and so on. The dipole term in the multipole expansion is

$$\Phi_{dipole} = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r'})r'}{r^2}(\sin\alpha)dv'.$$

Since the angle between $\boldsymbol{r}$ and $\boldsymbol{r'}$ is $(\frac{\pi}{2}-\alpha)$ then $\boldsymbol{e_r}\cdot\boldsymbol{r'} = r'\cos(\frac{\pi}{2}-\alpha) = r'\sin\alpha$ and so we may write the dipole term in a coordinate-independent manner as

$$\Phi_{dipole} = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r'})}{r^2}\,\boldsymbol{e_r}\cdot\boldsymbol{r'}dv' = \frac{1}{4\pi\epsilon_0}\frac{\boldsymbol{e_r}\cdot\boldsymbol{p}}{r^2}$$

where

$$\boldsymbol{p} = \int \rho(\boldsymbol{r'})\boldsymbol{r'}dv'$$

is called the dipole moment of a charge distribution. If instead of a volume charge density ρ we had a surface charge density σ or a linear charge density λ then we would defined the dipole moment in terms of a surface integral $\int \boldsymbol{r'}\sigma(\boldsymbol{r'})da$ or a line integral $\int \boldsymbol{r'}\lambda(\boldsymbol{r'})dl$, while for a collection of point charges the dipole moment is just the sum $\sum_i \boldsymbol{r_i'}q_i$. Consequently for the dipole of example 4.5, the dipole moment is

$$\boldsymbol{p} = q\boldsymbol{r_+'} - q\boldsymbol{r_-'} = q\begin{pmatrix}0\\0\\\frac{d}{2}\end{pmatrix} - q\begin{pmatrix}0\\0\\-\frac{d}{2}\end{pmatrix} = qd\,\boldsymbol{e_z}.$$

The dipole moment for a single dipole points along the direction from the negative charge to to the positive charge in the dipole and has length qd where q is the magnitude of the single charges used to construct the dipole and d the distance between those charges. It is worth remembering that we determined the form of the dipole potential by an approximation $r \gg d$,

so that the exact potential for the dipole will have quadrupole and higher pole contributions. The exact potential ϕ_{dipole} is called the potential for an ideal dipole[10].

Now that we have learned that the dipole potential is only the exact potential for an idealised dipole we may, nervously, wonder whether the monopole term

$$\Phi_{monopole} = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r'})}{r}dv'$$

is the exact potential for a point charge for a suitable charge density ρ. For a charge q located at the origin then $\rho(\boldsymbol{r'}) = q\delta^3(\boldsymbol{r'})$ and so then $\Phi_{monopole} = \frac{1}{4\pi\epsilon_0}\frac{q}{r}$, which is the exact potential for a point charge q located at the origin. What happens for a point charge q located somewhere other than the origin? One way to think about this is to commence with the potential for a point charge q located at the origin and apply a translation to the function. If instead of a charge q located at the origin we wished to consider the potential for a charge at $\boldsymbol{r'}$ we would find

$$\Phi(\boldsymbol{r} - \boldsymbol{r'}) = \frac{1}{4\pi\epsilon_0}\frac{q}{r - r'}$$

We may find $\Phi(\boldsymbol{r} - \boldsymbol{r'})$ via the Taylor expansion

$$\Phi(\boldsymbol{r} - \boldsymbol{r'}) = \Phi(\boldsymbol{r}) - \boldsymbol{r'}\cdot\boldsymbol{\nabla}\Phi(\boldsymbol{r}) + \frac{1}{2}(\boldsymbol{r'}\cdot\boldsymbol{\nabla})^2\Phi(\boldsymbol{r}) + \ldots$$

then

$$-\boldsymbol{r'}\cdot\boldsymbol{\nabla}\Phi(\boldsymbol{r}) = -\boldsymbol{r'}\cdot\Big(-\frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\boldsymbol{e_r}\Big) = \frac{1}{4\pi\epsilon_0}\frac{qr'}{r^2}\sin\alpha$$

$$\frac{1}{2}(\boldsymbol{r'}\cdot\boldsymbol{\nabla})^2\Phi(\boldsymbol{r}) = \frac{1}{2}\boldsymbol{r'}\cdot\boldsymbol{\nabla}\Big(-\frac{1}{4\pi\epsilon_0}\frac{qr'}{r^2}\sin\alpha\Big) = \frac{1}{4\pi\epsilon_0}\Big(\frac{qr'^2(3\sin^2\alpha - 1)}{2r^3}\Big)$$

To compute the second term above one must first realise that α is dependent upon $\boldsymbol{r}$. In Cartesian coordinates $\boldsymbol{r}\cdot\boldsymbol{r'} = xx' + yy' + zz' = rr'\sin\alpha$ so that

$$\begin{aligned}\frac{1}{2}(\boldsymbol{r'}\cdot\boldsymbol{\nabla})^2\Phi(\boldsymbol{r}) &= \frac{q}{2}\frac{1}{4\pi\epsilon_0}\boldsymbol{r'}\cdot\boldsymbol{\nabla}\Big(\frac{-(xx' + yy' + zz')}{r^3}\Big)\\ &= \frac{q}{2}\frac{1}{4\pi\epsilon_0}\begin{pmatrix}x'\\ y'\\ z'\end{pmatrix}\cdot\begin{pmatrix}\frac{3x(xx'+yy'+zz')}{r^5} - \frac{x'}{r^3}\\ \frac{3y(xx'+yy'+zz')}{r^5} - \frac{y'}{r^3}\\ \frac{3z(xx'+yy'+zz')}{r^5} - \frac{z'}{r^3}\end{pmatrix}\\ &= \frac{1}{4\pi\epsilon_0}\Big(\frac{qr'^2(3\sin^2\alpha - 1)}{2r^3}\Big).\end{aligned}$$

[10]This consists of taking the limit in which $d \to 0$, $q \to \infty$ such that $p = qd$ is constant. An ideal dipole sits at a point and has zero charge made out of two infinite contributions!

Therefore

$$\Phi(\boldsymbol{r}-\boldsymbol{r'}) = \frac{1}{4\pi\epsilon_0}\frac{q}{r}\Big(1+\frac{r'}{r}\sin\alpha+\frac{r'^2}{r^2}\Big(\frac{(1-3\sin^2\alpha)}{2}\Big)+\dots\Big)$$

which, in retrospect, we realise we could have anticipated due to our earlier computation in equation (4.9).

We may conclude that for the electric monopole, just as for the dipole, there is a perfect monopole which is a point charge located at the origin, whose potential contains a single monopole term, otherwise the potential for a point charge located away from the origin has a non-trivial multipole expansion. This leads to the delightfully absurd idea that we might model a point charge, not at the origin, by an infinite number of ideal multipoles located at the origin. Earlier we saw how we might approximate any distribution of charges by a sum of multipoles at the origin, now we see how we might understand the potential for a single charge away from the origin by an infinite number multipole charges at the origin.

5. Magnetostatics

In this chapter we will study magnetic forces due to magnetic fields which do not change, the static magnetic fields leads to this subject being known as magnetostatics. Our aim will be to investigate aspects of the mathematical description of magnetostatics following the same steps, or at least as close as we can manage, as we have analysed electrostatics. The reason for this is that there will be many familiar, but complementary, ideas in magnetostatics to those we have met in our study of electrostatics. Furthermore one may anticipate our study of electromagnetism, in which electrical forces and magnetic forces will be two aspects of one underlying system of dynamics.

5.1 The Lorentz Force Law

The analogue of Coulomb's law for electrostatics is the magnetic part of the Lorentz force law:

$$\boldsymbol{F}_{magnetic} = Q(\boldsymbol{v} \times \boldsymbol{B})$$

which is the force on a test charge Q, moving with velocity v in a magnetic field $\boldsymbol{B}$.

Comment(s).

- *The size of force due to a magnetic field is* $|F_{magnetic}| = |Q||\boldsymbol{B}||\boldsymbol{v}| \sin\alpha$ *where* α *is the angle between* $\boldsymbol{v}$ *and* $\boldsymbol{B}$*. We see that a magnetic field* $\boldsymbol{B}$ *produces no force on a static test charge* Q*. Furthermore the size of the force depends only on the component of the velocity field perpendicular to* $\boldsymbol{B}$ *i.e.* $|v| \sin\alpha$*. Furthermore if a test charge* Q *has a velocity parallel to a magnetic field then* $\alpha = 0$ *and the test charge feels no force due to the magnetic field. The appearance of* $\boldsymbol{v}$ *has the possibility of misleading us: in electrostatics all charges are static, which produces a static electric field. In magnetostatics the static field is not related to static electric charges, as we will se it is in fact derived from moving electric charges - the static in "magnetostatics" refers to the unchanging magnetic field that we will study, and not static charges.*

- *The direction of the force on a moving charge Q due to magnetic field is perpendicular to both the test charge's velocity $\boldsymbol{v}$ and the magnetic field $\boldsymbol{B}$. Indeed, the cross-product dictates precisely in which direction a charge Q experiences a force*[1]*. One of the most interesting consequences of the force law is that the direction of the force depends upon the sign of the charge. Consequently the charge of a particle may be determined from seeing which direction it is deflected in when moving through a magnetic field - the tracks in a bubble chamber give a practical way to do this.*

- *As $\boldsymbol{F}_{magnetic}$ is perpendicular to $\boldsymbol{v}$, then the force due to a magnetic field does not do any work, as $\boldsymbol{F}_{magnetic} \cdot d\boldsymbol{l} = \boldsymbol{F}_{magnetic} \cdot \boldsymbol{v} dt = 0$, since $\boldsymbol{F}_{magnetic}$ is perpendicular to $\boldsymbol{v}$. Magnetic forces do no work, they cannot accelerate a charged particle or slow it down, but they can alter the direction in which the test charge moves.*

By summing the force on a test charge Q due to an electric field $\boldsymbol{E}$ (Coulomb's law) and the force on the same test charge due to a magnetic field $\boldsymbol{B}$ we find the Lorentz force law[2]:

$$\boldsymbol{F} = Q\big(\boldsymbol{E} + (\boldsymbol{v} \times \boldsymbol{B})\big). \tag{5.1}$$

Since we are working with S.I. units where the force is measured in Newtons, then the magnetic field is measured in units called Tesla[3]. A one Tesla magnetic field is a very strong magnetic field, so it is frequently preferable to measure the magnetic field using the Gaussian unit, called the Gauss which is much smaller: 1 Tesla = 10,000 Gauss. The Earth's magnetic field ranges between 0.3 and 0.6 Gauss (0.00003-0.00006 Tesla), a small bar magnet produces around 100 Gauss (or 0.01 Tesla) and within a sunspot the magnetic field may be around 1500 Gauss (or 0.15 Tesla).

[1] One may use a right-hand-rule to derive the direction of the force due to a magnetic field on a test charge moving with velocity $\boldsymbol{v}$: Use, or imagine, the thumb, index and middle fingers of a right hand, and orient them so that the thumb aligns points along the positive x-axis, the index finger points along the positive y-axis and the middle finger points along the positive z-axis. These three fingers are held in this relative position in the following. The right-hand-rule for $\boldsymbol{F}_{magnetic} = Q(\boldsymbol{v} \times \boldsymbol{B})$ implies that the thumb is then aligned with direction of the component of $\boldsymbol{v}$ perpendicular to $\boldsymbol{B}$, the index finger is then aligned with direction of $\boldsymbol{B}$ and the resulting direction of the middle finger is the direction of the force.

[2] Named after Hendrik Lorentz, who wrote it down explicitly in 1895 even though it is implicitly included in Maxwell's earlier work (of 1865) and was also discovered explicitly before Lorentz by Oliver Heaviside in 1889.

[3] Which are equivalent to Newton-seconds per Coulomb-metre, or in the base units of the S.I. system kilograms per Ampère-second. It happens that an older unit the Weber per metre-squared, is also equal to 1 Tesla, and sometimes used.

5.2 Currents

In our investigation of electrostatics we were able to think about point charges or electric monopoles as sources of the electric field. Experiments demonstrate that the origin of constant magnetic fields are steady currents. A current is a charge in motion: imagine, for example, placing a loop of conducting material in an electric field arranged so as to drive the charges in the conductor perpetually around the loop producing a current. A battery will do just as well of course. In a one-dimensional conducting wire the current is defined to be the amount of charge passing a specified point on the wire in unit time, its units are called the Ampère, and 1 Ampère = 1 Coulomb per second.. Consequently if a line charge density λ is moving throughout a wire at speed v then in time δt $\lambda v \delta t$ charge passes any given point on the wire, so that the size of the current is $I = \lambda v$. However since the charge is confined to along the wire then we may speak of the velocity of the charge and consider the current in a wire to be the vector

$$\boldsymbol{I} = \lambda \boldsymbol{v}.$$

One oddity of the standard conventions is that the current measures the direction of movement of positive charge, hence in a metal wire, where the negatively charged electrons move, the current vector points in the opposite direction to the motion of electrons.

In the situation where there is a steady flow of charge in the wire we may compute the force on the wire due to an external magnetic field. In this case the charge Q in the Lorentz force law is the moving charge in the wire, which we know has density λ per unit length, so to find the total charge we must integrate over the length of the wire, whose shape will define the path in a line integral

$$\begin{aligned} \boldsymbol{F}_{magnetic} &= \int_P \lambda dl \, (\boldsymbol{v} \times \boldsymbol{B}) \\ &= \int_P dl \, (\boldsymbol{I} \times \boldsymbol{B}) \\ &= \int_P I \, (d\boldsymbol{l} \times \boldsymbol{B}) \end{aligned}$$

where in the last line we reasoned that the current is in the direction of the wire, so that both $Id\boldsymbol{l} = dl\boldsymbol{I}$, where $I = |\boldsymbol{I}|$.

Example 5.1. *Prove that the total force on a closed loop of wire carrying a current I (of constant magnitude) due to a constant magnetic field $\boldsymbol{B}$ is zero.*

We have

$$\boldsymbol{F} = \oint_P I(d\boldsymbol{l} \times \boldsymbol{B}).$$

The key to simplifying this problem is to notice what is varying under the integral, since I is constant and the magnetic field $\boldsymbol{B}$ is constant then we can write

$$\boldsymbol{F} = I\Big(\oint_P d\boldsymbol{l}\Big) \times \boldsymbol{B} = I(\boldsymbol{0}) \times \boldsymbol{B} = \boldsymbol{0}.$$

Where $\oint_P d\boldsymbol{l} = \boldsymbol{0}$ because we are integrating a total derivative over a closed loop (i.e. $\int_{\boldsymbol{a}}^{\boldsymbol{b}} d\boldsymbol{l} = \boldsymbol{l}(\boldsymbol{b}) - \boldsymbol{l}(\boldsymbol{a})$ but for a closed path then $\boldsymbol{a} = \boldsymbol{b}$). Note that it is the total force on the wire which is zero, but it is entirely possible that the force may be distributed in such a way as to cause the loop of wire to rotate.

For a surface charge of density σ in motion we define a surface current density by

$$\boldsymbol{K} = \sigma \boldsymbol{v}$$

then the force on the flowing current on a surface in a magnetic field $\boldsymbol{B}$ is

$$\boldsymbol{F} = \int_S \sigma da(\boldsymbol{v} \times \boldsymbol{B}) = \int_S (\boldsymbol{K} \times \boldsymbol{B}) da.$$

Similarly for a charge density ρ moving with velocity $\boldsymbol{v}$ through a volume we may define a volume current density

$$\boldsymbol{J} = \rho \boldsymbol{v}$$

so that the force on a volume through which a steady current is flowing is

$$\boldsymbol{F} = \int_V \rho dv(\boldsymbol{v} \times \boldsymbol{B}) = \int_V (\boldsymbol{J} \times \boldsymbol{B}) dv.$$

The volume current density is a vector that represents how much charge flows through a surface of unit area at right angles to the flow of charge. Consequently if we consider the infinitesimal area vector $d\boldsymbol{a}$ on a surface, which we recall is a vector normal to the surface then the amount of charge flowing across that surface area element in unit time is

$$\boldsymbol{J} \cdot d\boldsymbol{a} = \rho v \cos(\theta) da$$

we see that in time dt the charge crossing the surface is $\rho(v\cos(\theta)dt)da$, the role of the dot product is to ensure that we just consider the component of the velocity normal to the surface when measuring how much charge passes in time dt. Therefore the total charge flowing through a surface per unit time is the flux of $\boldsymbol{J}$ the volume current density. For a closed surface S the total flow of charge across the surface per unit time is the amount of charge leaving the volume enclosed by the surface[4], which we may express as

$$\oint_S \boldsymbol{J} \cdot d\boldsymbol{a} = -\frac{dQ_S}{dt}$$

[4]We are presuming that charge is conserved, so that charge flowing out of a volume corresponds to a decrease in the charge within the volume.

where we have first considered the total change in the charge Q_S in the surface in a small finite time δt, before diving through by the small time and taking its limit to zero to obtain the derivative on the left-hand-side. Since charge is leaving the enclosed volume we have written this expression with a minus sign. Alternatively, we may also compute the charge contained within the volume V using the volume charge density:

$$Q_S = \int_V \rho dv$$

Substituting this into the previous integral we have

$$\oint_S \boldsymbol{J} \cdot d\boldsymbol{a} = -\frac{d}{dt}\Big(\int_V \rho dv \Big)$$

and by using Green's theorem on the integral on the left-hand-side gives us

$$\int_V (\boldsymbol{\nabla} \cdot \boldsymbol{J})\, dv = -\int_V \frac{\partial \rho}{\partial t}\, dv.$$

The partial derivatives $\frac{\partial \rho}{\partial t}$ appear because ρ is, in general a function of the coordinates as well as of time. Now from the equation above we deduce

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} = -\frac{\partial \rho}{\partial t}. \tag{5.2}$$

This is an equation describing the conservation of charge, it is called the continuity equation.

5.3 The Biot-Savart Law and Ampère's Law

Stationary charges are the source of the static electric field, but it is steady currents which are the source of the static magnetic field. Together these give the conditions under which we study both electrostatics and magnetostatics:

$$\frac{d\rho}{dt} = 0$$
$$\frac{d\boldsymbol{J}}{dt} = \boldsymbol{0}.$$

What is a steady current? It is a current that does not change over time, a flow of charge that continues unchanged forever. How is this consistent with $\frac{d\rho}{dt} = 0$? In a physical wire, the net charge is zero, but the current is carried by the motion of electrons - even though the electrons are in motion the net charge in the wire (coming from the electrons and the protons) remains constant. So long as there is not a build up of charge in the wire then a steady

current is consistent with a constant charge density. One immediate consequence of working with magnetostatics is that

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} = 0.$$

This is interesting, our experience with electrostatics tells us that the field lines of $\boldsymbol{J}$ cannot begin or end on any local points (compare this with $\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}$ where the electric field can begin or end on any charges which give rise to a non-zero value of ρ).

What is the physical rule that governs how steady currents in wires produce a static magnetic field? The rule is known as the Biot-Savart law which, like the Coulomb law and the Lorentz force law, we will take as an axiom:

$$\begin{aligned} \boldsymbol{B}(\boldsymbol{r}) &= \frac{\mu_0}{4\pi} \int_P \frac{\boldsymbol{I} \times \boldsymbol{e_s}}{s^2} dl' \\ &= \frac{\mu_0}{4\pi} I \int_P \frac{d\boldsymbol{l'} \times \boldsymbol{e_s}}{s^2} \end{aligned} \tag{5.3}$$

where we have taken the constant I out of the integral using $\boldsymbol{I} = I d\boldsymbol{l'}$ for a current in a wire, P is the path along the wire, and $\boldsymbol{s} = \boldsymbol{r} - \boldsymbol{r'}$ is the displacement vector between a test charge Q at $\boldsymbol{r}$ and an element of current carrying wire at $\boldsymbol{r'}$. The coefficient $\mu_0 = 4\pi \times 10^{-7}$ Newtons-per-Ampère-squared is called the permeability of free space, and appears in the equation in order to assist with definition of units (in this case the Ampère).

Example 5.2. *Find the magnetic field a fixed distance ρ from an infinite straight wire on the z-axis carrying a steady current I travelling in the direction $\boldsymbol{e_z}$:*

Due to the symmetry of the problem we realise that the magnetic field must be invariant under a translation along the z-axis. So, for simplicity, we consider the magnetic field at a point in the plane $z = 0$, i.e. we take $\boldsymbol{r} = \begin{pmatrix} \rho\cos\theta \\ \rho\sin\theta \\ 0 \end{pmatrix}$. *Now $\boldsymbol{r'}$ is the location of the infinitesimal element of current-carrying wire, so it lies on the z-axis, i.e.* $\boldsymbol{r'} = \begin{pmatrix} 0 \\ 0 \\ z' \end{pmatrix}$. *Therefore* $\boldsymbol{s} = \boldsymbol{r} - \boldsymbol{r'} = \begin{pmatrix} \rho\cos\theta \\ \rho\sin\theta \\ -z' \end{pmatrix}$ *and* $\boldsymbol{e_s} = \frac{1}{\sqrt{\rho^2 + (z')^2}} \begin{pmatrix} \rho\cos\theta \\ \rho\sin\theta \\ -z' \end{pmatrix}$. *Since $d\boldsymbol{l'} = dz'\,\boldsymbol{e_z}$ then we have*

$$\begin{aligned} d\boldsymbol{l'} \times \boldsymbol{s} &= (dz'\,\boldsymbol{e_z}) \times (\rho\cos\theta\,\boldsymbol{e_x} + \rho\sin\theta\,\boldsymbol{e_y} - z'\,\boldsymbol{e_z}) \\ &= \rho\cos\theta dz'\,\boldsymbol{e_y} - \rho\sin\theta dz'\,\boldsymbol{e_x} \\ &= \rho dz'\,\boldsymbol{e_\theta} \end{aligned}$$

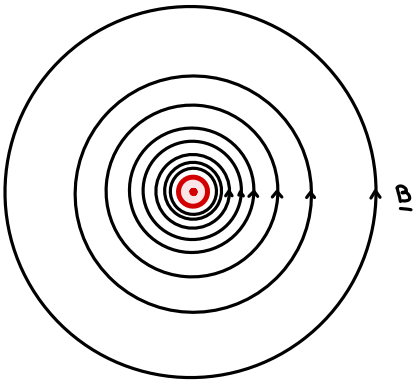
where in the last line we have substituted the angular basis vector $\boldsymbol{e_\theta} = (-\sin\theta)\,\boldsymbol{e_x} + (\cos\theta)\,\boldsymbol{e_y}$ *in cylindrical coordinates. Note that we have computed* $d\boldsymbol{l'} \times \boldsymbol{s}$ *rather than* $d\boldsymbol{l'} \times \boldsymbol{e_s}$ *as* $\boldsymbol{e_s} = \frac{\boldsymbol{s}}{s}$*, so we may swap* $\boldsymbol{s}$ *and* $\boldsymbol{e_s}$ *if we are interested in just the direction of* $\boldsymbol{B}$*, but when we work out* $|\boldsymbol{B}|$ *we must then take care to include an extra power of* s *in the denominator under the integral in the Biot-Savart law. The magnetic field due to the infinite wire points along the tangent to circles centred on the wire, so the magnetic field lines form circles around the wire which lie in planes perpendicular to the wire. It remains for us to compute the size of the magnetic field.*

$$\begin{aligned}
|\boldsymbol{B}(\boldsymbol{r})| &= \frac{\mu_0}{4\pi} I \int_{z'=-\infty}^{\infty} \frac{\rho dz'}{(\rho^2 + (z')^2)^{\frac{3}{2}}} \\
&= \frac{\mu_0}{4\pi} I \int_{\psi'=-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\cos\psi}{\rho} d\psi \\
&= \frac{\mu_0}{2\pi} \frac{I}{\rho}
\end{aligned}$$

where we have made the substitution $z' = \rho\tan\psi$*. Hence*

$$\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0 I}{2\pi\rho}\,\boldsymbol{e_\theta}$$

where $\boldsymbol{r}$ *is a point a distance* ρ *(radially) from the wire. The magnetic field lines in a plane that the wire cuts perpendicularly are concentric circles about the wire:*



Current coming out of the page.

There is a right-hand rule for working out the direction of the magnetic field about a current: imagine closing the fingers of your right hand with the thumb on the same hand pointing in the direction of $\boldsymbol{I}$*, then the magnetic field turns in the same direction as your fingers curl when the fist of your right-hand is closed.*

Example 5.3. *Consider two infinite parallel wires, both parallel to the* z*-axis, separated by a distance* r *which carry currents* I_1 *and* I_2 *respectively in the positive* z*-direction. Find the forces which act on a length* L *of the two wires.*

From the previous example we know that each current produces a magnetic field, the current I_1 produces a magnetic field $\boldsymbol{B}_1 = \frac{\mu_0 I_1}{2\pi r}\boldsymbol{e_\theta}$ on the second wire (a distance r away), while the current I_2 produces a magnetic field $\boldsymbol{B}_2 = \frac{\mu_0 I_2}{2\pi r}\boldsymbol{e_\theta}$, albeit in cylindrical coordinates centred about the second wire rather than the first, at the location of the first wire. Let us suppose the first wire sits on the z-axis, while the second wire is parallel to it but a distance r away. The first wire produces a magnetic field $\boldsymbol{B}_1 = \frac{\mu_0 I_1}{2\pi r}\boldsymbol{e_\theta}$ which exerts a force (due to the Lorentz force law) on the moving charge in the second wire. We will denote this force as $\boldsymbol{F}_{21}$, standing for the force on wire two due to the magnetic field produced by wire one, and we have, for a length L of the wire a force:

$$\begin{aligned}\boldsymbol{F}_{21} &= \int_P I_2(d\boldsymbol{l}_2' \times \boldsymbol{B}_1) \\ &= \int_{z'=0}^{L} I_2(dz'\,\boldsymbol{e_z} \times \frac{\mu_0 I_1}{2\pi r}\boldsymbol{e_\theta}) \\ &= -\frac{\mu_0 I_1 I_2 L}{2\pi r}\boldsymbol{e_\rho}.\end{aligned}$$

where we have used the identity $\boldsymbol{e_\theta} \times \boldsymbol{e_z} = \boldsymbol{e_\rho}$ computed in section C.1. Notice that these are the cylindrical coordinates set up around the first wire, and the force acts in the $-\boldsymbol{e_\rho}$ direction on wire two, i.e. it is drawn towards wire one.

Carrying out a similar computation for the force on the current in a length of the first wire we find (for a set of cylindrical coordinates now centred about the second wire)

$$\begin{aligned}\boldsymbol{F}_{12} &= \int_P I_1(d\boldsymbol{l}_1' \times \boldsymbol{B}_2) \\ &= \int_{z'=0}^{L} I_1(dz'\,\boldsymbol{e_z} \times \frac{\mu_0 I_2}{2\pi r}\boldsymbol{e_\theta}) \\ &= -\frac{\mu_0 I_1 I_2 L}{2\pi r}\boldsymbol{e_\rho}.\end{aligned}$$

The force again (as we would have expected by symmetry) acts radially inwards, but now we have constructed our coordinates about the location of wire two. Our shift in coordinate frames obscures the fact that the forces are equal and opposite, the two wires are attracted to each other (as in both frames the forces points directly towards the other wire). If current in one of the wires flowed in the opposite direction, introducing a single minus sign into our computations, then the two wires repel each other.

5.3.1 The Divergence and Curl of $\boldsymbol{B}$

Consider the magnetic field around the infinitely long, straight wire along the z-axis carrying a current $\boldsymbol{I} = I\,\boldsymbol{e_z}$ which we found in the example above to be

$$\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0 I}{2\pi r}\,\boldsymbol{e_\theta}$$

where r is the radial distance from the axis on which the current lies. Since the field is constructed of concentric circles we can anticipate that its will have a non-zero circulation density and hence the curl of $\boldsymbol{B}$ will not be zero.

From Stokes' theorem we have

$$\int_S (\boldsymbol{\nabla} \times \boldsymbol{B}) \cdot d\boldsymbol{a} = \oint_P \boldsymbol{B} \cdot d\boldsymbol{l}$$

so for the magnetic field around the current-carrying wire along a path P which follows a concentric circle of constant magnetic field strength we find

$$\begin{aligned}
\int_S (\boldsymbol{\nabla} \times \boldsymbol{B}) \cdot d\boldsymbol{a} &= \oint_P \boldsymbol{B} \cdot d\boldsymbol{l} \\
&= \int_{\theta=0}^{2\pi} (\frac{\mu_0 I}{2\pi r}\,\boldsymbol{e_\theta}) \cdot (r d\theta\,\boldsymbol{e_\theta}) \\
&= \frac{\mu_0 I}{2\pi} \int_{\theta=0}^{2\pi} d\theta \\
&= \mu_0 I.
\end{aligned}$$

Now consider the case where the path is no longer a concentric circle, just some closed loop encircling the wire in the plane perpendicular to the wire. The difference for the non-circular path is that r is now a function of θ and varies under the integral. The path is P is parameterised by θ as $\begin{pmatrix} r(\theta)\cos\theta \\ r(\theta)\sin\theta \\ z \end{pmatrix}$, so that

$$d\boldsymbol{l} = \begin{pmatrix} r'(\theta)\cos\theta\, d\theta - r(\theta)\sin\theta\, d\theta \\ r'(\theta)\sin\theta d\theta + r(\theta)\cos\theta\, d\theta \\ 0 \end{pmatrix} = r'(\theta) d\theta\,\boldsymbol{e_\rho} + r(\theta) d\theta\,\boldsymbol{e_\theta}.$$

Hence

$$\boldsymbol{B}(r(\theta)) \cdot d\boldsymbol{l} = (\frac{\mu_0 I}{2\pi r(\theta)}\,\boldsymbol{e_\theta}) \cdot (r'(\theta) d\theta\,\boldsymbol{e_\rho} + r(\theta) d\theta\,\boldsymbol{e_\theta}) = \frac{\mu_0 I}{2\pi} d\theta$$

as for the earlier calculation for the circular loop. So the result is unchanged when the path is any closed loop, in the plane perpendicular to the current, about the current-carrying wire.

Finally suppose that the plane of the closed path is no longer perpendicular to the current, so that the path is given by the vectors $\begin{pmatrix} r(\theta)\cos\theta \\ r(\theta)\sin\theta \\ z(\theta) \end{pmatrix}$ where $\theta \in [0, 2\pi]$ parameterises the closed path. Then

$$d\boldsymbol{l} = \begin{pmatrix} r'(\theta)\cos\theta\, d\theta - r(\theta)\sin\theta\, d\theta \\ r'(\theta)\sin\theta d\theta + r(\theta)\cos\theta\, d\theta \\ z'(\theta)\, d\theta \end{pmatrix} = r'(\theta)d\theta\, \boldsymbol{e_\rho} + r(\theta)d\theta\, \boldsymbol{e_\theta} + z'(\theta)\, d\theta\, \boldsymbol{e_z}.$$

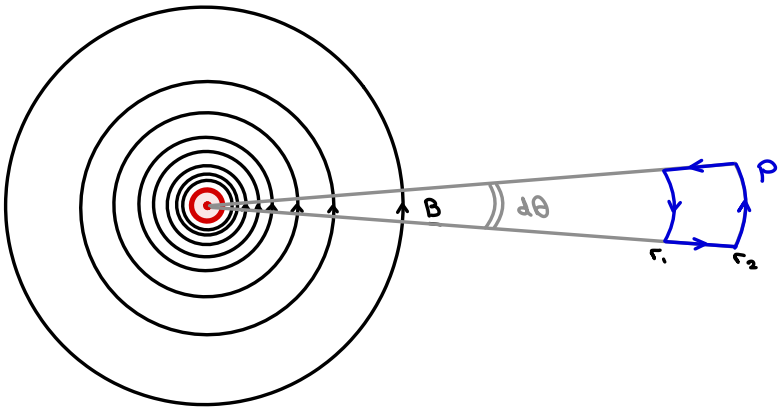
So we find again that

$$\boldsymbol{B}(r(\theta)) \cdot d\boldsymbol{l} = \frac{\mu_0 I}{2\pi} d\theta.$$

Hence the integral $\oint_P \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 I$ for any closed path, in a plane, about a long straight current-carrying wire. This is, of course, reminiscent of Gauss's law in electrostatics - albeit for a line integral and for a single steady current. However we may now add current-carrying wires within the closed path, and add the resulting magnetic fields to find that

$$\oint_P \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 I_P$$

where I_P is the total current carried by wires crossing the area S within the closed path P. Any magnetic field produced by a current that does not cross the area S surrounded by the closed path P contributes zero to the line integral. Consider the closed loop P drawn in the plane about a current-carrying wire but not enclosing the wire:



Current coming out of the page.

The line integral $\oint_P \boldsymbol{B} \cdot d\boldsymbol{l}$ has no contributions from the two parts of the path P parallel to the radial direction $\boldsymbol{e_\rho}$, on the inner segment at r_1 we have $d\boldsymbol{l}_1 = -r_1 d\theta\, \boldsymbol{e_\theta}$ so that $\boldsymbol{B} \cdot d\boldsymbol{l}_1 = -\frac{I\mu_0}{2\pi} d\theta$

and on the outer segment at r_2 we have $d\boldsymbol{l}_2 = r_2 d\theta\, \boldsymbol{e_\theta}$ so that $\boldsymbol{B} \cdot d\boldsymbol{l}_2 = \frac{I\mu_0}{2\pi} d\theta$. Hence the contribution to $\boldsymbol{B} \cdot d\boldsymbol{l}$ from the closed loop P not surrounding the current is zero.

Now suppose the location of the current is described by a current density $\boldsymbol{J}$ then the total current I_P in a region bounded by P is the flux of J through S

$$I_P = \int_S \boldsymbol{J} \cdot d\boldsymbol{a}$$

hence, returning to Stokes' theorem, we have

$$\begin{aligned} \int_S (\boldsymbol{\nabla} \times \boldsymbol{B}) \cdot d\boldsymbol{a} &= \oint_P \boldsymbol{B} \cdot d\boldsymbol{l} \\ &= \mu_0 I_P \\ &= \mu_0 \int_S \boldsymbol{J} \cdot d\boldsymbol{a} \end{aligned}$$

from which we deduce that

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J}.$$

Our result is valid for straight line currents but we can show that it is generally valid. Let us restate the Biot-Savart law in terms of a volume current density $\boldsymbol{J}$:

$$\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0}{4\pi} \int_V \frac{\boldsymbol{J}(\boldsymbol{r'}) \times \boldsymbol{e_s}}{s^2} dv' \tag{5.4}$$

Recalling that $\boldsymbol{\nabla}(\frac{1}{s}) = -\frac{\boldsymbol{s}}{s^3} = -\frac{1}{s^2}\boldsymbol{e_s}$ hence we may develop the Biot-Savart law as

$$\begin{aligned} \boldsymbol{B}(\boldsymbol{r}) &= -\frac{\mu_0}{4\pi} \int_V \boldsymbol{J}(\boldsymbol{r'}) \times \boldsymbol{\nabla}(\frac{1}{s})\, dv' \\ &= \frac{\mu_0}{4\pi} \boldsymbol{\nabla} \times \int_V \frac{\boldsymbol{J}(\boldsymbol{r'})}{s}\, dv' \end{aligned}$$

where in the second line we have used the product law for the curl:

$$\boldsymbol{\nabla} \times (\frac{1}{s} J(\boldsymbol{r'})) = \boldsymbol{\nabla}(\frac{1}{s}) \times \boldsymbol{J}(\boldsymbol{r'}) + \frac{1}{s} \boldsymbol{\nabla} \times \boldsymbol{J}(\boldsymbol{r'}) = -\boldsymbol{J}(\boldsymbol{r'}) \times \boldsymbol{\nabla}(\frac{1}{s})$$

where $\boldsymbol{\nabla} \times \boldsymbol{J}(\boldsymbol{r'}) = \boldsymbol{0}$ because $\boldsymbol{\nabla}$ is a function of $\boldsymbol{r}$ and not $\boldsymbol{r'}$. Since the magnetic field can be written as the curl of $\frac{\mu_0}{4\pi} \int_V \frac{\boldsymbol{J}(\boldsymbol{r'})}{s}\, dv'$ then we realise that

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

as $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{v}) = 0$ for any vector field $\boldsymbol{v}$. Comparing this equation to Gauss's Law $\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}$ we realise it implies that the magnetic field does not have any local sinks or sources, or, in

other words, there are no magnetic monopoles[5]. Magnetic fields are not sourced by points of magnetic charge (if it was then the magnetic field would not satisfy $\boldsymbol{\nabla}\cdot\boldsymbol{B}=0$ at that point).

The curl of $\boldsymbol{B}$ becomes

$$\boldsymbol{\nabla}\times\boldsymbol{B}(\boldsymbol{r})=\boldsymbol{\nabla}\times\Big(\boldsymbol{\nabla}\times\frac{\mu_0}{4\pi}\int_V\frac{\boldsymbol{J}(\boldsymbol{r'})}{s}\,dv'\Big)$$

Now we have met second vector derivatives of this form before, in equation (2.11), we found that $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{v})=\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{v})-\boldsymbol{\nabla}^2\boldsymbol{v}$ for a vector field $\boldsymbol{v}$. Consequently,

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{B}(\boldsymbol{r})&=\boldsymbol{\nabla}\Big(\boldsymbol{\nabla}\cdot(\frac{\mu_0}{4\pi}\int_V\frac{\boldsymbol{J}(\boldsymbol{r'})}{s}\,dv')\Big)-\boldsymbol{\nabla}^2\Big(\frac{\mu_0}{4\pi}\int_V\frac{\boldsymbol{J}(\boldsymbol{r'})}{s}\,dv'\Big)\\&=\boldsymbol{\nabla}\Big(\frac{\mu_0}{4\pi}\int_V\boldsymbol{J}(\boldsymbol{r'})\cdot\boldsymbol{\nabla}(\frac{1}{s})\,dv'\Big)-\frac{\mu_0}{4\pi}\int_V\boldsymbol{J}(\boldsymbol{r'})\boldsymbol{\nabla}^2(\frac{1}{s})\,dv'\\&=-\boldsymbol{\nabla}\Big(\frac{\mu_0}{4\pi}\int_V\boldsymbol{J}(\boldsymbol{r'})\cdot\boldsymbol{\nabla}'(\frac{1}{s})\,dv'\Big)+\mu_0\int_V\boldsymbol{J}(\boldsymbol{r'})\delta^3(\boldsymbol{s})\,dv'\end{aligned}$$

where in the second term we have made use of $\boldsymbol{\nabla}^2\Big(\frac{1}{s}\Big)=-4\pi\delta^3(\boldsymbol{s})$ found in equation (2.16). But what have we done in the first term? We have introduced $\boldsymbol{\nabla}'=\begin{pmatrix}\frac{\partial}{\partial x'}\\\frac{\partial}{\partial y'}\\\frac{\partial}{\partial z'}\end{pmatrix}$ and then since $\boldsymbol{s}=\boldsymbol{r}-\boldsymbol{r'}$ then $\boldsymbol{\nabla}(s)=-\boldsymbol{\nabla}'(s)$ and so also $\boldsymbol{\nabla}(\frac{1}{s})=-\boldsymbol{\nabla}'(\frac{1}{s})$. The reason we have made this change is so that we can now use integration by parts under the integral to develop a term $\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{r'})$ which we know is equal to zero. Since $\boldsymbol{\nabla}'(\boldsymbol{J}\frac{1}{s})=\boldsymbol{\nabla}'(\frac{1}{s})\cdot\boldsymbol{J}+\frac{1}{s}\boldsymbol{\nabla}'\cdot\boldsymbol{J}$ we have

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{B}(\boldsymbol{r})&=-\boldsymbol{\nabla}\Big(\frac{\mu_0}{4\pi}\int_V\Big(\boldsymbol{\nabla}'\big(\boldsymbol{J}(\boldsymbol{r'})\frac{1}{s}\big)-\frac{1}{s}\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{r'})\Big)\,dv'\Big)+\mu_0\boldsymbol{J}(\boldsymbol{r})\\&=\mu_0\boldsymbol{J}(\boldsymbol{r})\end{aligned}$$

where we have presumed that the term $\int_V\Big(\boldsymbol{\nabla}'\big(\boldsymbol{J}(\boldsymbol{r'})\frac{1}{s}\big)dv'$ is zero, in doing so we are making the assumption that the current density $\boldsymbol{J}$ is zero or constant at infinity, in which case it is suppressed by the term $\frac{1}{s}$ in the integral. For the vanishing of the other term under the integral we have applied our earlier observation that the current density has zero divergence, $\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{r'})=0$.

[5]This is a prediction that rests on the validity of the Biot-Savart law, but has, so far, been confirmed by our knowledge of the universe: no magnetic monopole has ever been detected. One of the motivating reasons for proposing cosmological inflation was to explain why magnetic monopoles have not been observed. On the other hand, Dirac showed that even if just one magnetic monopole existed in the universe then from quantum mechanical arguments one could use its existence to explain the quantisation of electric charge.

We now have

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \tag{5.5}$$

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J} \tag{5.6}$$

of which equation (5.6) is called Ampère's law. We saw earlier that we could use Stokes' theorem to derive Ampère's law in integral form:

$$\oint_P \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 I_P$$

where I_P is the total current cutting through the surface bounded by P. These equations (in differential form) are the magnetostatic counterparts of our electrostatic equations

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}$$

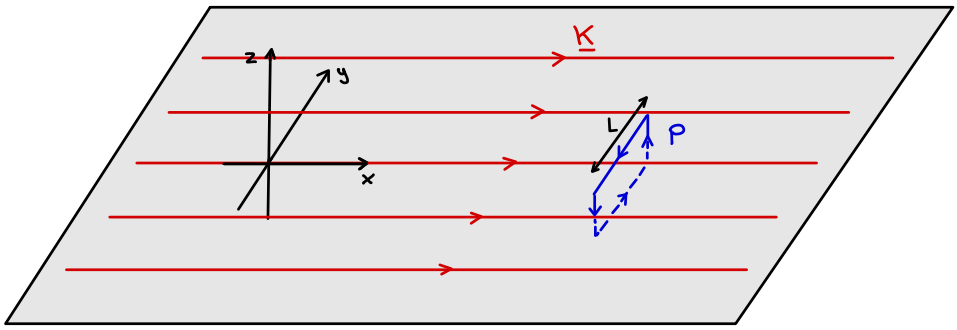
$$\boldsymbol{\nabla} \times \boldsymbol{E} = \boldsymbol{0}$$

and will allow us to find the magnetic field in settings which are sufficiently symmetric, although, in this setting, instead of Gaussian surfaces we will refer to Ampèrian surfaces. Furthermore the equations will allow us to construct a magnetic potential function, although for the magnetic field the potential will be a vector field.

Example 5.4. *Find the magnetic field of a uniform surface current $\boldsymbol{K} = K\,\boldsymbol{e_x}$ flowing in the plane $z = 0$.*

First, let us think about possible magnetic field lines. They cannot end or begin anywhere other than at infinity as $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$ so the field cannot begin or end on the plane. This idea rules out the possibility of a constant magnetic field perpendicular to the plane (such a field would have field lines ending on the plane). We may also use our knowledge about the magnetic field around current in a long, straight wire: the plane of current is really a sum of such wires, and so we may expect that the magnetic field is parallel to the plane but perpendicular to the direction of flow of current, and points in opposite directions either side of the plane. These ideas suggest a path on which we may compute $\oint_P \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 I_P$ so that $\boldsymbol{B}$ is constant on the integration path. We consider the path, called an Ampèrian loop, which forms a rectangle which

is perpendicular to the direction of the current and surrounds some of the current in the plane:



The Ampèrian loop has edges parallel to the y-axis of length L. Now using Ampère's law we have,

$$\begin{aligned}\oint_P \boldsymbol{B}\cdot d\boldsymbol{l} &= 2BL\\ &= \mu_0 I_P\\ &= \mu_0 KL\end{aligned}$$

Now our arguments about the direction of the magnetic field mean that it is pointing in the $-\boldsymbol{e_y}$ direction for $z>0$ and in the $\boldsymbol{e_y}$ direction for $z<0$. Therefore

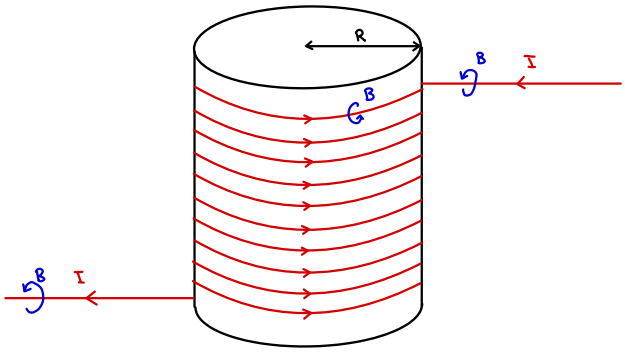
$$\boldsymbol{B}=\begin{cases}-\frac{\mu_0 K}{2}\boldsymbol{e_y} & z>0,\\ \frac{\mu_0 K}{2}\boldsymbol{e_y} & z<0.\end{cases}$$

As for the electric field over the infinite charged plane, the field is independent of the distance from the plane.

Example 5.5. *Find the magnetic field of an infinitely long, tightly wound, solenoid*[6] *consisting of n turns of wire per unit length on a cylinder of radius R, where each loop of wire carries a steady current I. We will assume that each loop of wire forms a circle of radius R centred on*

[6] A solenoid is a helical coil of wire which you may imagine as loops of wire wrapping a cylinder.

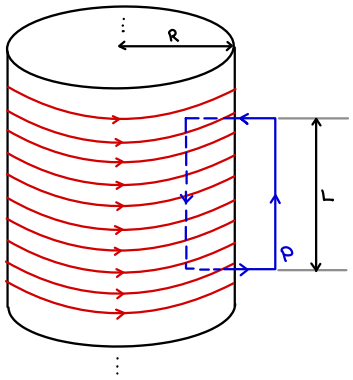
the axis of the cylinder. We show here a sketch of a finite solenoid:



In order to find a path on which the integral of $\boldsymbol{B}$ *in Ampère's law simplifies, we must first find the direction of* $\boldsymbol{B}$*. Now around each loop of current on the solenoid we may use a right-hand-rule (with thumb pointing along the direction of current) so that our fingers curl in the same direction that* $\boldsymbol{B}$ *circulates about the current. For the solenoid depicted above, aligned with the z-axis, the magnetic field must sum to point along the direction* $\boldsymbol{e_z}$ *inside the solenoid, but it must point in the* $-\boldsymbol{e_z}$ *direction outside the solenoid. Consequently we will pick an Ampèrian loop which is a rectangle with two edges parallel to* $\boldsymbol{e_z}$ *and the remaining edges parallel to* $\boldsymbol{e_\rho}$*. First consider such a path that bounds an area outside the solenoid (through which no current cuts) then we can apply Ampère's law to deduce that* $\oint_P \boldsymbol{B}\cdot d\boldsymbol{l} = 0$ *outside the solenoid is zero (as there is no current there), which implies that the magnetic field is constant outside the solenoid*[7]*. Later in the lectures on we will show that, incredibly, it is in fact zero. Now consider an Ampèrian loop , which cuts the loops of wire on the solenoid. The rectangle*

[7]For a rectangular loop of height L in the plane spanned $\boldsymbol{e_\rho}$ and $\boldsymbol{e_z}$ for some fixed θ, then $\oint_P \boldsymbol{B}\cdot d\boldsymbol{l} = \int_{z=0}^{L}(-B(\rho_1)\,\boldsymbol{e_z})\cdot(-dz\,\boldsymbol{e_z}) + \int_{z=0}^{L}(-B(\rho_2)\,\boldsymbol{e_z})\cdot(dz\,\boldsymbol{e_z}) = (B(\rho_1)-B(\rho_2))L = 0$, where the edges of the Ampèrian rectangle parallel to $\boldsymbol{e_z}$ are located at ρ_1 and ρ_2. Hence $B(\rho_1) = B(\rho_2)$.

for this integration path will have length L in the $\boldsymbol{e_z}$ direction:



Now we have argued that

$$\boldsymbol{B}(r) = \begin{cases} B_{Inside}\,\boldsymbol{e_z} & r < R \\ -B_{Outside}\,\boldsymbol{e_z} & r > R \end{cases}$$

where $B_{Outside}$ is a constant. Then

$$\oint_P \boldsymbol{B} \cdot d\boldsymbol{l} = -B_{Inside}L - B_{Outside}L = \mu_0 I_P = -\mu_0 nLI.$$

Consequently

$$\boldsymbol{B}(r) = \begin{cases} (\mu_0 nI - B_{Outside})\,\boldsymbol{e_z} & r < R \\ -B_{Outside}\,\boldsymbol{e_z} & r > R \end{cases}$$

Now notice that if $B_{Outside} > \mu_0 nI$ then the magnetic field points in the wrong direction inside the solenoid, i.e. it must be the case that $B_{Outside} < \mu_0 nI$. Since $B_{Outside} \geq 0$ and because the current may be arbitrarily small we deduce that $B_{Outside} = 0$ giving

$$\boldsymbol{B}(r) = \begin{cases} \mu_0 nI\,\boldsymbol{e_z} & r < R \\ \boldsymbol{0} & r > R. \end{cases}$$

5.4 The Magnetic Vector Potential

The equation $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$ is automatically satisfied if

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

where $\boldsymbol{A}$ is called the magnetic vector potential. In terms of the magnetic vector potential Ampère's law becomes

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{B} &= \boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A})\\ &= \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A})-\boldsymbol{\nabla}^2\boldsymbol{A}\\ &= \mu_o\boldsymbol{J}\end{aligned}$$

which does not directly yield an equation of the simplicity of Poisson's equation in electrostatics. However we still have some room for manoeuvre, as the magnetic vector potential is not uniquely defined but is only defined up to a gradient. Consider the curl of $\boldsymbol{A}+\boldsymbol{\nabla}\Lambda$ where Λ is a scalar function,

$$\boldsymbol{\nabla}\times(\boldsymbol{A}+\boldsymbol{\nabla}\Lambda)=\boldsymbol{\nabla}\times\boldsymbol{A}=\boldsymbol{B}$$

as $\boldsymbol{\nabla}\times\boldsymbol{\nabla}(\Lambda)=\boldsymbol{0}$ for any scalar function Λ. Let us write $\boldsymbol{A}=\boldsymbol{A_0}+\boldsymbol{\nabla}\Lambda$, and now we may wonder if we could choose Λ such that $\boldsymbol{\nabla}\cdot\boldsymbol{A}=0$. Choosing Λ in this way so as to make $\boldsymbol{\nabla}\cdot\boldsymbol{A}=0$ is called fixing the gauge, and this choice is known as the Coulomb gauge. Being able to do this would imply

$$0=\boldsymbol{\nabla}\cdot\boldsymbol{A}=\boldsymbol{\nabla}\cdot\boldsymbol{A_0}+\boldsymbol{\nabla}^2\Lambda \implies \boldsymbol{\nabla}^2\Lambda=-\boldsymbol{\nabla}\cdot\boldsymbol{A_0}.$$

Our experience of Poisson's equation and the Green's function for the Laplacian means we can solve the equation $\boldsymbol{\nabla}^2\Lambda=-\boldsymbol{\nabla}\cdot\boldsymbol{A_0}$ readily[8] to find

$$\Lambda=\frac{1}{4\pi}\int_V\frac{\boldsymbol{\nabla}\cdot\boldsymbol{A_0}}{s}dv'$$

Taking this value for Λ we then have

$$\boldsymbol{\nabla}^2\boldsymbol{A}=-\mu_0\boldsymbol{J}.$$

The equation above is the magnetostatic analogue of Poisson's equation and its solution is

$$\boldsymbol{A}(\boldsymbol{r})=\frac{\mu_0}{4\pi}\int_V\frac{\boldsymbol{J}(\boldsymbol{r'})}{s}dv'.$$

Example 5.6. *Find the vector potential of the magnetic field $\boldsymbol{B}=B\,\boldsymbol{e_z}$ where B is a constant.*

We seek $\boldsymbol{A}$ such that $\boldsymbol{\nabla}\times\boldsymbol{A}=\boldsymbol{B}=B\,\boldsymbol{e_z}$, so $\boldsymbol{A}$ circulates about the z-axis, i.e. $\boldsymbol{A}=A_\theta\,\boldsymbol{e_\theta}$ in cylindrical coordinates, as then $\boldsymbol{\nabla}\times\boldsymbol{A}=-\frac{\partial A_\theta}{\partial z}\boldsymbol{e_\rho}+\left(\frac{1}{\rho}\frac{\partial(\rho A_\theta)}{\partial\rho}\right)\boldsymbol{e_z}=B\,\boldsymbol{e_z}$. So $\frac{\partial A_\theta}{\partial z}=0$ and therefore $\int_0^{\rho_0}\frac{\partial(\rho A_\theta)}{\partial\rho}d\rho=\frac{1}{2}B\rho_0^2$, hence $A_\theta(\rho_0)=\frac{1}{2}B\rho_0$. Therefore $\boldsymbol{A}=\frac{1}{2}B\rho\,\boldsymbol{e_\theta}$, or in

[8] It is as if we were finding the electric potential for the charge density $\epsilon_0\boldsymbol{\nabla}\cdot\boldsymbol{A_0}$.

Cartesian coordinates $\boldsymbol{A} = \frac{1}{2}B(-y\,\boldsymbol{e_x} + x\,\boldsymbol{e_y})$. *Now we may add any gradient* $\boldsymbol{\nabla}\Lambda$ *to* $\boldsymbol{A}$ *to obtain the same magnetic field* $\boldsymbol{B}$ *for example if* $\Lambda = \frac{1}{2}x^2$ *then* $\boldsymbol{\nabla}\Lambda = x\,\boldsymbol{e_x}$ *and so we expect that* $\boldsymbol{A}+\boldsymbol{\nabla}\Lambda = (-\frac{1}{2}By+x)\,\boldsymbol{e_x}+\frac{1}{2}Bx\,\boldsymbol{e_y}$ *gives the same magnetic field* $\boldsymbol{B}$*, which we confirm by directly computing the curl. Interestingly we can choose* Λ *so as to eliminate either the* x*-component or the* y*-component of* $\boldsymbol{A}$*. For example by choosing* $\Lambda = \frac{1}{2}Bxy$*, then* $\boldsymbol{\nabla}\Lambda = \frac{1}{2}B(y\,\boldsymbol{e_x} + x\,\boldsymbol{e_y})$ *so that* $\boldsymbol{A} + \boldsymbol{\nabla}\Lambda = Bx\,\boldsymbol{e_y}$ *while* $\boldsymbol{A} - \boldsymbol{\nabla}\Lambda = -By\,\boldsymbol{e_x}$*. The form of neither of these magnetic potentials immediately gives us the impression that they are circulating about the* z*-axis and generating* $\boldsymbol{B} = B\,\boldsymbol{e_z}$*. So one must take care with drawing quick inferences from the vector magnetic potential, some reflection is always a good idea.*

Example 5.7. *Find the vector potential of the magnetic field produced by an infinitely long wire along the* z*-axis carrying steady current* $\boldsymbol{I} = I\,\boldsymbol{e_z}$*.*

We have found the magnetic field for an infinitely long wire carrying a current $\boldsymbol{I} = I\,\boldsymbol{e_z}$ *earlier, in terms of cylindrical coordinates it is* $\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0 I}{2\pi\rho}\,\boldsymbol{e_\theta}$*. Now from*

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int_V \frac{\boldsymbol{J}(\boldsymbol{r'})}{s}dv'.$$

with $\boldsymbol{J} = I\,\boldsymbol{e_z}\delta(x)\delta(y)$ *we immediately see that the* x *and* y*-components of* $\boldsymbol{A}$ *are zero, so we have* $\boldsymbol{A} = A_z\,\boldsymbol{e_z}$ *(in cylindrical coordinates as well as Cartesian coordinates) and we must determine* A_z*. Now computing the curl in cylindrical coordinates (using equation* (C.2)*) we have*

$$\boldsymbol{\nabla}\times\boldsymbol{A} = \frac{1}{\rho}\frac{\partial A_z}{\partial\theta}\,\boldsymbol{e_\rho} - \frac{\partial A_z}{\partial\rho}\,\boldsymbol{e_\theta} = \boldsymbol{B} = \frac{\mu_0 I}{2\pi\rho}\,\boldsymbol{e_\theta}.$$

Comparing the coefficients of the basis vectors we deduce that $\frac{\partial A_z}{\partial\theta} = 0$ *and* $-\frac{\partial A_z}{\partial\rho} = \frac{\mu_0 I}{2\pi\rho}$ *from which we deduce that*

$$\boldsymbol{A} = -\frac{\mu_0 I}{2\pi}\ln(\rho)\,\boldsymbol{e_z}.$$

Example 5.8. *Find the magnetic vector potential of an infinitely long solenoid along the* z*-axis carrying steady current* $\boldsymbol{I} = I\,\boldsymbol{e_\theta}$ *in the wire with* n *turns of wire per unit length and of radius* R

The vector potential $\boldsymbol{A}$ *is parallel to the current density vector and hence we expect to be able to find* $\boldsymbol{A}$ *with only its* θ *component non-zero. Now,*

$$\oint_P \boldsymbol{A}\cdot d\boldsymbol{l} = \int_S (\boldsymbol{\nabla}\times\boldsymbol{A})\cdot d\boldsymbol{a} = \int_S \boldsymbol{B}\cdot d\boldsymbol{a}$$

which is the flux of $\boldsymbol{B}$ *through the surface* S *bounded by the loop* P*. Let* P *be a circle of fixed radius* ρ *in the plane* $z = 0$ *then, since by symmetry* A *must be constant over* P*, we have*

$$\oint_P \boldsymbol{A} \cdot d\boldsymbol{l} = \int_{\theta=0}^{2\pi} (A_\theta \, \boldsymbol{e_\theta}) \cdot (\rho d\theta \, \boldsymbol{e_\theta}) = 2\pi\rho A_\theta.$$

For $\rho < R$ *we have found that* $\boldsymbol{B} = \mu_0 nI \, \boldsymbol{e_z}$ *hence*

$$\int_S \boldsymbol{B} \cdot d\boldsymbol{a} = \int_{r=0}^{\rho} \int_{\theta=0}^{2\pi} (\mu_0 nI \, \boldsymbol{e_z})(r \, dr \, d\theta \, \boldsymbol{e_z}) = \mu_0 nI\pi\rho^2.$$

Which gives us

$$\boldsymbol{A} = \frac{\mu_0 nI\rho}{2} \boldsymbol{e_\theta} \qquad \text{for } \rho < R.$$

Outside the solenoid ($\rho \geq R$) *we compute*

$$\int_S \boldsymbol{B} \cdot d\boldsymbol{a} = \int_{r=0}^{R} \int_{\theta=0}^{2\pi} (\mu_0 nI \, \boldsymbol{e_z})(r \, dr \, d\theta \, \boldsymbol{e_z}) = \mu_0 nI\pi R^2.$$

as the magnetic flux is only non-zero within the solenoid (of radius R*). Hence*

$$\boldsymbol{A} = \frac{\mu_0 nIR^2}{2\rho} \boldsymbol{e_\theta} \qquad \text{for } \rho \geq R.$$

The vector potential outside the solenoid is non-zero, and decreases as $\frac{1}{\rho}$*, even though the magnetic field is zero in the same region.*

5.5 The Multipole Expansion for the Vector Potential

Just as we were able to expand the electric potential in powers of $\frac{1}{r}$ to find the multipole expansion of Φ, we are also able to find the multipole expansion for the magnetic vector potential for a given loop of current. In working out the multipole expansion the crucial step was deducing that

$$\frac{1}{s} = \frac{1}{r} \sum_{n=0}^{\infty} \left(\frac{r'}{r}\right)^n P_n(\sin\alpha).$$

Since the magnetic potential for a steady current I running in a loop P is

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0 I}{4\pi} \oint_P \frac{1}{s} d\boldsymbol{l}'$$

we may use the expansion of $\frac{1}{s}$ in terms of $\frac{1}{r}$ to find the multipole expansion of the magnetic potential

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0 I}{4\pi} \frac{1}{r} \oint_P \sum_{n=0}^{\infty} \left(\frac{r'}{r}\right)^n P_n(\sin\alpha) d\boldsymbol{l}'.$$

As for the electric potential the term proportional to $\frac{1}{r}$ is called the magnetic monopole term, the term proportional to $\frac{1}{r^2}$ is called the magnetic dipole term and so on. Note that the magnetic monopole term is, in fact, trivial:

$$\boldsymbol{A}_{Monopole}(\boldsymbol{r}) = \frac{\mu_0 I}{4\pi}\frac{1}{r}\oint_P d\boldsymbol{l}' = \boldsymbol{0}.$$

Consequently the leading term in any magnetic multipole expansion is the dipole term:

$$\begin{aligned}\boldsymbol{A}_{Dipole}(\boldsymbol{r}) &= \frac{\mu_0 I}{4\pi}\frac{1}{r}\oint_P \frac{r'}{r}\sin\alpha\, d\boldsymbol{l}' \\ &= \frac{\mu_0 I}{4\pi}\frac{1}{r^2}\oint_P \boldsymbol{e_r}\cdot\boldsymbol{r'}\, d\boldsymbol{l}'\end{aligned}$$

A useful identity that is simple to prove (see problem sheet VII, problem 35) is

$$\oint_P \boldsymbol{e_r}\cdot\boldsymbol{r'}\, d\boldsymbol{l}' = -\,\boldsymbol{e_r}\times\int_S d\boldsymbol{a}'$$

and it allows us to rewrite the dipole term for the vector potential as

$$\boldsymbol{A}_{Dipole}(\boldsymbol{r}) = \frac{\mu_0 I}{4\pi}\frac{1}{r^2}\int_S d\boldsymbol{a}'\times\boldsymbol{e_r} \equiv \frac{\mu_0}{4\pi}\frac{\boldsymbol{m}\times\boldsymbol{e_r}}{r^2}$$

where $\boldsymbol{m}$ is called the magnetic dipole moment and is defined by

$$\boldsymbol{m} = I\int_S d\boldsymbol{a} = I\boldsymbol{a}$$

where $\boldsymbol{a}$ is the vector area of the surface S bounded by P, its direction being the integral of the normal directions to the surface (its sign is picked out by the right-hand-rule, so that if your fingers curl around in the direction of the current loop then your thumb points in the direction of $\boldsymbol{a}$). For example, if P is the circle of radius R in the xy-plane with current running in the $\boldsymbol{e_\theta}$ direction then $d\boldsymbol{a} = r\,dr\,d\theta\,\boldsymbol{e_z}$, then $\boldsymbol{a} = \pi R^2\,\boldsymbol{e_z}$ and the magnetic dipole moment for this loop of current is $\boldsymbol{m} = I\pi R^2\,\boldsymbol{e_z}$.

6. Electrodynamics

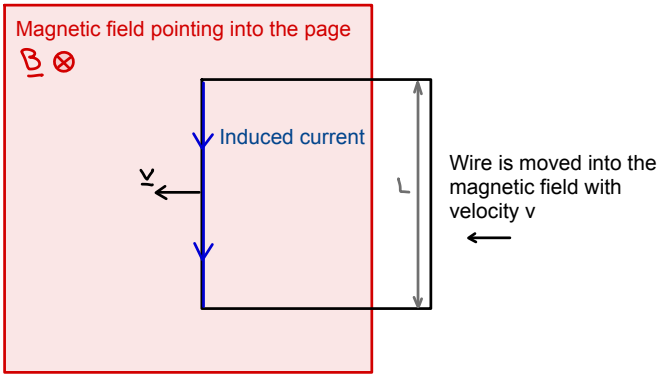
In introducing magnetic forces we stated that a magnetic field is produced by moving charges in the form of a current, and in the Lorentz force law we have seen that a moving electric charge feels a force due to a magnetic field. These ideas lead the way towards unifying electrostatics and magnetostatics in the single subject of electrodynamics, which is be the study of changing electric and magnetic fields.

6.1 Maxwell's Equations

We will now develop two of the equations of electrostatics and magnetostatics; the resulting set of equations will be Maxwell's equations.

6.1.1 Electromotive Forces and Faraday's Law

Since a charge Q moving with velocity $\boldsymbol{v}$ in a magnetic field $\boldsymbol{B}$ is subject to a force $\boldsymbol{F}_{magnetic} = Q(\boldsymbol{v} \times \boldsymbol{B})$, we may wonder what effect a magnetic field has on the charge in a conductor. In particular let us consider a wire, made of a conductor, but not carrying any current, which is moved into a magnetic field as depicted below:



The motion of the wire as a whole means that the charge (which is moving with the wire, but not around the wire) has a velocity $\boldsymbol{v} = -v\,\boldsymbol{e_x}$, which in the magnetic field $\boldsymbol{B} = -B\,\boldsymbol{e_z}$ gives rise to a force

$$\boldsymbol{F}_{magnetic} = Q((-v\,\boldsymbol{e_x}) \times (-B\,\boldsymbol{e_z})) = -QvB\,\boldsymbol{e_y}$$

on each charge Q, which for the charges in the part of the wire aligned with the y-axis cause them to accelerate along that section of the wire, and so inducing a current. The charges in the parts of the wire parallel to the x-axis cannot move in the y-direction, as they are constrained to remain within the wire[1]. If the loop of wire was moving in the opposite direction (i.e. if $\boldsymbol{v} = v\,\boldsymbol{e_x}$) then the induced current would move in a clockwise direction around the loop of wire. The motion of the wire is relative to the magnetic field, so we may equally imagine that the current is induced in a stationary wire due to a moving magnetic field. We are now in the realm of electrodynamics.

The Lorentz force does work on the charges in the left-hand part of the wire, but only in the part of the wire parallel to the $\boldsymbol{e_y}$ direction in the magnetic field, along the rest of the circuit the charges are not able to move (as the Lorentz force is perpendicular to the wire) or there is no Lorentz force, because the wire is outside of the magnetic field. Upon encountering a force, it is natural to ask what work it does, in this case to discover how much work a moving magnetic field can do as it brings charge in the wire into motion. Working with the force per unit charge we compute the work done around the loop of wire[2]

$$\oint_P \frac{1}{Q}\boldsymbol{F}_{magnetic} \cdot d\boldsymbol{l} = vBL \equiv \mathcal{E}.$$

where we have introduced

$$\mathcal{E} = \oint_P \frac{1}{Q}\boldsymbol{F}_{magnetic} \cdot d\boldsymbol{l} \tag{6.1}$$

which is called the electromotive force. In considering the work done by the force per unit charge we sidestep problems such as: which charge are we focussing on and does it make a circuit of the wire? Furthermore we observe that the force per unit charge is precisely the meaning of the electric field in electrostatics, so whatever quantity it is that we are considering it will be a quantity that is similar to the electric field.

[1]However the moving charge leaves a charge deficit which has an electric field which causes charge to move around the loop. We will think about this idea again when we think about how batteries cause current to flow in a wire.

[2]We are integrating in an anticlockwise direction around the loop (as viewed from looking down the z-axis towards the xy-plane), so on the left edge of the wire - where the Lorentz force acts we integrate in the negative y-direction e.g. $\int_L^0 -vBdy = vBL$.

The electromotive force is a quantity whose origin is in the changing magnetic field. For example, suppose that there had been an additional background static electric field $\boldsymbol{E}$ then the total force per unit charge would have a contribution due to the magnetic field and the electric field, but the computation above would remain unchanged:

$$\oint_P (\frac{1}{Q}\boldsymbol{F}_{magnetic} + \boldsymbol{E}) \cdot d\boldsymbol{l} = vBL \equiv \mathcal{E}$$

as $\oint_P \boldsymbol{E} \cdot d\boldsymbol{l} = 0$. The quantity we have computed, then, is entirely due to a changing magnetic field - it is an electrodynamic quantity. The electromotive force for the simple loop above can be expressed in another way: $-\frac{dx}{dt}BL = \mathcal{E}$ so that $\int BLdx = -\int \mathcal{E}dt$. The integral on the left-hand-side is the flux of $\boldsymbol{B}$ over the area S bounded by the loop of wire which we denote

$$\Phi_B = \int_S \boldsymbol{B} \cdot d\boldsymbol{a}$$

for a surface S.[3] Hence the equation for the electromotive force in our rectangular loop of wire becomes,

$$\frac{d\Phi_B}{dt} = -\mathcal{E} \tag{6.2}$$

note that, for the wire moving to the left, the flux of $\boldsymbol{B}$ is actually decreasing (this is because of the fact that the normal vector $d\boldsymbol{a} \propto \boldsymbol{e_z}$ is in the opposite direction to $\boldsymbol{B} \propto -\boldsymbol{e_z}$), and that $\mathcal{E} = vBL > 0$ so that the signs in the equation above are consistent.

While we have deduced equation (6.2) for our rectangular wire, we can show that it holds for a loop of wire of any shape. Consider the motion of a wire of velocity $\boldsymbol{v}$

$$\begin{aligned}\mathcal{E} &= \oint_P \frac{1}{Q}\boldsymbol{F}_{magnetic} \cdot d\boldsymbol{l} \\ &= \oint_P (\boldsymbol{v} \times \boldsymbol{B}) \cdot d\boldsymbol{l} \\ &= \oint_P (d\boldsymbol{l} \times \boldsymbol{v}) \cdot \boldsymbol{B}.\end{aligned}$$

[3]Note that we have taken $d\boldsymbol{a}$ to be in the $\boldsymbol{e_z}$ direction, i.e. $d\boldsymbol{a} = \begin{pmatrix} 0 \\ 0 \\ dxdy \end{pmatrix}$, the opposite direction to $\boldsymbol{B}$ - this is consistent with the right-hand-rule for the anticlockwise path around the wire in the line integral. Notice, for later, that for the contributing part of the line integral (i.e. along the leading left edge of the circuit, for an anticlockwise path) $d\boldsymbol{l} = \begin{pmatrix} 0 \\ -dy \\ 0 \end{pmatrix}$ and $\boldsymbol{v} = \begin{pmatrix} -v \\ 0 \\ 0 \end{pmatrix}$, then $(\boldsymbol{v} \times d\boldsymbol{l})dt = d\boldsymbol{a}$, consistent with the right-hand-rule convention for picking the overall sign of the normal to an open surface.

Now an element of wire $d\boldsymbol{l}$ moves a distance $\boldsymbol{v}dt$ in a time dt and sweeps out an infinitesimal area $|(d\boldsymbol{l} \times \boldsymbol{v})dt| = |d\boldsymbol{a}|$ and we define[4] $(\boldsymbol{v} \times d\boldsymbol{l})dt = d\boldsymbol{a}$ hence

$$\begin{aligned}\int \mathcal{E}dt &= \int \oint_P (d\boldsymbol{l} \times \boldsymbol{v}) \cdot \boldsymbol{B}dt \\ &= -\int_S d\boldsymbol{a} \cdot \boldsymbol{B} \\ &= -\Phi_B\end{aligned}$$

where the integral over dt has combined with the integral over $d\boldsymbol{l} \times \boldsymbol{v}$ to give a surface integral. We have again obtained the result that the electromotive force is equal to minus the flux of $\boldsymbol{B}$ through the surface S surrounded by the wire, but now the wire may be of any closed shape. There is a proliferation of minus signs in the above discussion which does not quite form a conspiracy. The guide-rails to keep hold of are that the normal vector should point in the direction indicated by the right-hand-rule applied around the path in the line integral (i.e. curl the fingers of a right hand in the direction of the path of the line integral, and the normal should be chosen to point in the direction of the thumb). We will introduce Lenz's law later, and the minus sign in $\mathcal{E} = -\frac{d\Phi_B}{dt}$ can be understood as encapsulating the idea that the current will flow in a direction to oppose the change in the flux of $\boldsymbol{B}$ through the circuit.

Notice that the electromotive force depends on the shape of the wire, it also depends on the relative velocity of the wire and the magnetic field, and it depends on the size of the magnetic field. We may conclude that a current will flow in a wire if the wire moves in the magnetic field, if the magnetic field moves with respect to the stationary wire, or if the magnitude of the magnetic field varies. These three ideas were first observed experimentally by Michael Faraday in 1831 and result in a modification to the laws of electrostatics that we have been working with so far, let us explain why. In the first of these ideas, where the charges in the wire have a velocity $\boldsymbol{v}$, we understand that the current is a result of the magnetic force term in the Lorentz force law. The second, where the magnetic field creates the same current when it is itself in motion, could not be determined directly from the Lorentz force law, where $\boldsymbol{v}$ is the velocity of the test charge (the charges in the wire) - however we are used to relative motion and the idea that we could view the second situation from a frame that moves with the magnetic field, and in that frame we would see the wire moving and then again apply the Lorentz force law to find the current is produced by the magnetic field in the moving conductor. The third case however, defies our interpretation. Certainly the flux of $\boldsymbol{B}$ through S changes if the magnitude

[4] As ever, we aim to be consistent with the right-hand rule: a line integral on an anticlockwise path in the xy-plane should give a normal in the $\boldsymbol{e_z}$ direction, e.g. $d\boldsymbol{l} = \begin{pmatrix} 0 \\ dy \\ 0 \end{pmatrix}$ and $\boldsymbol{v}dt = \begin{pmatrix} dx \\ 0 \\ 0 \end{pmatrix}$ gives $(\boldsymbol{v} \times d\boldsymbol{l})dt = dxdy\,\boldsymbol{e_z}$.

of $\boldsymbol{B}$ varies with time, or indeed if $\boldsymbol{B}$ changes in any way such that $\int_S \boldsymbol{B} \cdot d\boldsymbol{a} = \Phi_B$ does not remain constant. This situation induces an electromotive force as $\mathcal{E} = -\frac{d\Phi_B}{dt}$ and current in the wire, but without there being any relative velocity $\boldsymbol{v}$ between the wire and the magnetic field: this current cannot be explained as arising from the magnetic force in the Lorentz force law (as $\boldsymbol{v} = \boldsymbol{0}$). It must be that the electric force term in the Lorentz force law is responsible for this current when $\frac{d\boldsymbol{B}}{dt} \neq \boldsymbol{0}$: the changing magnetic field has produced an electric field. Via the flux of $\boldsymbol{B}$ we have one unified way to compute the current induced in the wire when $\boldsymbol{B}$ changes with time. In all three cases the change in Φ_B produces an electromotive force, and now we may equate these essentially two different origins of the current (1. the moving magnetic field, and 2. our deduction that there must be an induced electric field when $\frac{d\boldsymbol{B}}{dt} \neq \boldsymbol{0}$) and discover a modification to our laws of electrostatics. In the case where there is an induced electric field $\boldsymbol{E}$ driving the charge around the circuit we have

$$\mathcal{E} = \oint_P \boldsymbol{E} \cdot d\boldsymbol{l} = \int_S (\boldsymbol{\nabla} \times \boldsymbol{E}) \cdot d\boldsymbol{a}$$

where we have made use of Stokes' theorem, while we also know that

$$\mathcal{E} = -\frac{d\Phi_B}{dt} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot d\boldsymbol{a}$$

where we have used the definition of the magnetic flux in the second line[5]. Now following in Faraday's steps we may equate these two ways of thinking about the electromotive force in order to deduce the electric field in the loop of wire. Consequently we have

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}$$

which is called Faraday's Law of Induction in differential form. For the case of a static magnetic field the equation reduces to one of our equations of electrostatics, as expected. This is the first of the modifications of our equations from electrostatics and magnetostatics due to dynamical fields.

We could summarise Faraday's equation in the oversimplified statement that a changing magnetic field induces an electric field, a better summary is to say that an electromotive force is produced around a loop P whenever the magnetic flux through the surface S, bounded by P, changes. Looking back at Faraday's three observations, we understand that in all cases the magnetic flux through S for the loop of wire P is changing, however when the loop is moving

[5]Why has the total derivative become a partial derivative under the integral? First we observe that, under the integral, only $\boldsymbol{B}$ might be a function of time. The surface S is fixed so $d\boldsymbol{a}$ is not a function of time. The magnetic field $\boldsymbol{B}$ is generally a function of position and time - hence the use of the partial derivative with respect to time.

then it is directly the magnetic field $\boldsymbol{B}$ that produces the current, when it is the field $\boldsymbol{B}$ which is moving or just changing in time, then there is an induced electric field which is responsible for the driving the charge around the circuit. We are also left with a small puzzle to think about: when the circuit moves it is the magnetic field that produces the current, while when we consider the magnetic field to be moving but the circuit stationary then it is the electric field which produces the current. It seems unsatisfactory that the frame of reference is so pivotal to deciding which field is causing the current to flow. We will revisit this idea.

Faraday's law contains sufficient signs so as to render it a little difficult to work out the direction of the induced electromotive force in a circuit (and hence the current), there is a useful rule, called Lenz's law, for working this out: 'the induced current will flow in the direction such that the magnetic flux it induces aims to reduce the initial magnetic flux'. The electromotive force induces a current whose own magnetic field attempts to counter the increasing or decreasing magnetic flux: change in flux is resisted!

Thoughts about the magnetic field induced by a current, lead us to another novel idea: consider a current flowing in a conductor, and now let us allow the current to change - this results in a changing magnetic field which then induces a current in the conductor itself. This new current is such that it opposes the changing magnetic field and so opposes the change in the current in the wire (which is causing the change in the magnetic field...), the net result is that a change in current in a wire is impeded, there is an inertia associated with currents.

6.1.2 Correcting Ampère's Law

The differential form of Ampère's law is

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J}.$$

Now while $\boldsymbol{\nabla} \times \boldsymbol{B}$ is mathematically divergence-free, the current density does not have zero divergence. Its divergence is given by the continuity equation which we derived in equation (5.2) which states that

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} = -\frac{\partial \rho}{\partial t}.$$

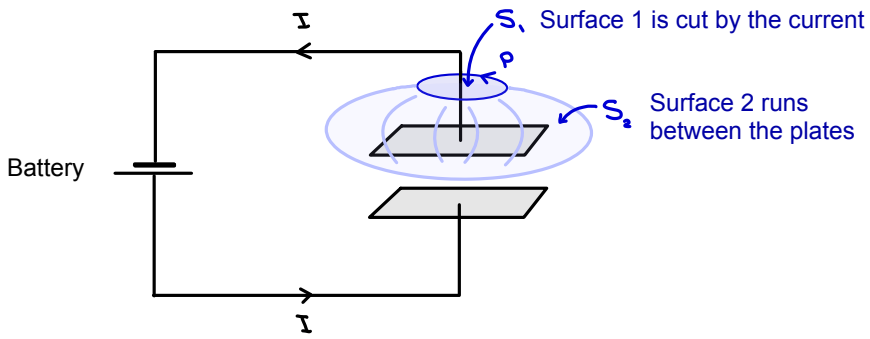
When we were studying the static fields, one of our assumptions was that $\frac{\partial \rho}{\partial t} = 0$, this was sufficient to guarantee that both sides of Ampère's law have zero divergence. In electrodynamics we will not assume that charge distributions do not change with time, so we will need to amend Ampère's law so as to guarantee that both sides are zero. The continuity equation tells us that the right-hand-side fails to be divergence-less by a term $\mu_0 \frac{\partial \rho}{\partial t}$. Since we know that $\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}$

we may posit that

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J} + \epsilon_0 \mu_0 \frac{\partial \boldsymbol{E}}{\partial t}$$

because $\boldsymbol{\nabla} \cdot (\mu_0 \boldsymbol{J} + \epsilon_0 \mu_0 \frac{\partial \boldsymbol{E}}{\partial t}) = \mu_0 \boldsymbol{\nabla} \cdot \boldsymbol{J} + \epsilon_0 \mu_0 \frac{\partial}{\partial t}(\boldsymbol{\nabla} \cdot \boldsymbol{E}) = -\mu_0 \frac{\partial \rho}{\partial t} + \epsilon_0 \mu_0 \frac{\partial}{\partial t}\left(\frac{\rho}{\epsilon_0}\right) = 0$. Adding the term $\epsilon_0 \mu_0 \frac{\partial \boldsymbol{E}}{\partial t}$ to Ampère's law was one of Maxwell's key insights, and while he was motivated for several reasons the support he had for this new equation was initially entirely theoretical. We may summarise the new term as telling us that a change in the electric field induces a magnetic field. A few examples will help to convince us of the need for this extra term.

Example 6.1. *Consider a current I that is being driven around a loop of wire by a battery (see below for a discussion of how a battery does this), and now suppose that the wire is broken and two conducting parallel plates, separated by a small gap, are inserted in the circuit. Now consider the Ampèrian loop P shown below*



and consider the two different surfaces S_1 (which is cut by the current) and S_2 (which is not cut by the current but runs between the two parallel plates) which both have the boundary P. Explain how Maxwell's modification of Ampère's law guarantees that $\oint_P \boldsymbol{B} \cdot d\boldsymbol{l} = \int_{S_1} (\boldsymbol{\nabla} \times \boldsymbol{B}) \cdot d\boldsymbol{a} = \int_{S_2} (\boldsymbol{\nabla} \times \boldsymbol{B}) \cdot d\boldsymbol{a}$.

We have drawn a circuit diagram above that includes a battery and we will digress for a moment from our problem to give an idea about how a battery can generate a uniform current all the way around a wire. A battery delivers an electromotive force to the charges in a circuit once its terminals are connected to the circuit. The electromotive force is measured in volts, this is the voltage of the battery. When the battery's terminals are connected to a conducting wire a current flows through the battery as the electromotive force causes the mobile charges in the conducting wire to move (the electrons in a metal conductor). Now the wire may be very long or it may have an unusual shape, so initially it seems unlikely that setting up a flow of charge near the battery is enough to create a uniform current in the wire, we might instead

expect that the current may be large near the battery and small far away from the battery, or that a change of direction in the wire might slow down or speed up charge as it flows, leading to an irregular current. However suppose for some reason that the current was not uniform and charge built up at some part of the wire, then the accumulated charge would produce an electric field which would repel like charges from that point. Near the battery once the circuit is connected the electromotive force of the battery causes charge near the battery to move, that moving charge leaves behind a net charge which generates its own electric field so as to bring more charge in to replace the charge moved by the battery and so on around the circuit. The fact that the electric field is so strong, and its action so quick ensures that a steady current is almost immediately generated throughout the wire due to the presence of the battery.

Let us return to our example and consider the integral over surface S_1 which is cut by the current. This is the kind of problem we have worked with already when studying magnetostatics, and if we pick S_1 to be a disc of radius R is cut by the current then we know that

$$\begin{aligned}\oint_P \boldsymbol{B}\cdot d\boldsymbol{l} &= \int_{S_1} (\boldsymbol{\nabla}\times\boldsymbol{B})\cdot d\boldsymbol{a}\\ &= \mu_0 \int_{S_1} \boldsymbol{J}\cdot d\boldsymbol{a}\\ &= \mu_0 I\end{aligned}$$

we note that over this surface the electric field is constant: there is a steady current and we are in the realm of electrostatics.

The integral over the second surface S_2 runs over a surface that runs between the two parallel conducting plates, but which is not cut by the current. Consequently we realise that there will be no contribution to $\int_{S_2}(\boldsymbol{\nabla}\times\boldsymbol{B})\cdot d\boldsymbol{a}$ from the current density (since it is zero over the surface S_2). We may guess at this stage that the term $\epsilon_0\mu_0\frac{\partial \boldsymbol{E}}{\partial t}$ in Maxwell's correction of Ampére's law must not be zero, in order to guarantee that $\oint_P \boldsymbol{B}\cdot d\boldsymbol{l}$ takes the same value no matter which surface (with boundary P) that we pick. To understand this we will need to think a little about capacitors.

A capacitor is a pair of parallel conducting plates which carry equal but opposite charge on each plate. When the distance between the plates is small, and the area of each plate large, the electric field in a capacitor can be approximated by the electric field between two infinite charged planes of charge density $-\sigma$ and σ. We found the electric field of such a system in problem 12 of the tutorial problems: there is only a non-zero electric field between the two plates whose magnitude is $\frac{\sigma}{\epsilon_0}$ and points in the direction normal to the plates away from the positively charged plate and towards the plate with negative charge.

Returning to our circuit, the battery causes charge to build up on the conducting plates of equal and opposite size. The current causes the lower plate in our sketch of the circuit to become

positively charged (we recall that the direction of the current is the effective direction of motion of positive charge) and the top plate to become negatively charged and an electric field builds up between the plates pointing from the lower plate to the upper plate, whose magnitude grows as the current causes more charge to be deposited on the plates. By definition $\boldsymbol{I} = \lambda \boldsymbol{v}$ *where* λ *is the linear charge density in the wire, and* $\boldsymbol{v}$ *is the velocity of the charge, consequently the current causes a charge* $Q(t) = \int_0^t \lambda v \, dt$ *to build up in a time* t *on the positive plate, and* $-Q(t)$ *on the negative plate. If the plates have area* A *then the electric field between the plates is, approximately,* $\boldsymbol{E} = \frac{Q(t)}{A\epsilon_0}\hat{\boldsymbol{n}}$ *where* $\hat{\boldsymbol{n}}$ *is the unit vector normal to the plates and pointing from the positive to the negative plates. Now we have*

$$\begin{aligned}
\oint_P \boldsymbol{B} \cdot d\boldsymbol{l} &= \int_{S_2} (\boldsymbol{\nabla} \times \boldsymbol{B}) \cdot d\boldsymbol{a} \\
&= \int_{S_2} \epsilon_0 \mu_0 \frac{\partial \boldsymbol{E}}{\partial t} \cdot d\boldsymbol{a} \\
&= \int_A \epsilon_0 \mu_0 \frac{\partial}{\partial t} \big(\frac{Q(t)}{A\epsilon_0} \hat{\boldsymbol{n}} \big) \cdot d\boldsymbol{a} \\
&= \int_A \mu_0 \frac{\partial Q(t)}{\partial t} \big(\frac{1}{A} \hat{\boldsymbol{n}} \big) \cdot (da \, \hat{\boldsymbol{n}}) \\
&= \mu_0 \lambda v \frac{1}{A} \int_A da \\
&= \mu_0 I
\end{aligned}$$

where we note that because the plates are assumed to be close together and because the electric field is only changing between the plates, that the integral of $\frac{\partial \boldsymbol{E}}{\partial t}$ *is only non-zero over the area* A *of the plates, which is why the integral changes from being over* S_2 *to over* A*.*

Hence Maxwell's term which is added to Ampère's equation guarantees that the line integral of $\boldsymbol{B}$ *over* P *is independent of the surface chosen for this circuit.*

Example 6.2. *Consider a spherically symmetric radial distribution of current density* $\boldsymbol{J}(r)$*, of the same magnitude in every direction. Let us denote by* $Q(R)$ *the total charge contained within a sphere of radius* R *centred at the origin. Now consider a circular path* P *of small radius on the surface of a sphere about the origin, the radial currents cut this path so we may deduce that a magnetic field circulates about the path* P*. However if we chose a different path* P *we would argue that the magnetic field must circulate about this second choice of* P*. The more we think about it more we realise that there is no consistent choice of the direction of* $\boldsymbol{B}$ *- a calculation with a different path in the line integral will give a different result. How does Maxwell's modification of Ampère's law resolve this problem?*

In the description of the problem there was a focus on the current generating the circulation of the magnetic field, but we know that a changing electric field will also contribute to $\boldsymbol{\nabla} \times \boldsymbol{B}$*.*

In this problem the electric field is changing: current is carrying charge away from $Q(R)$ the total charge contained in the sphere of radius R. Charge is lost from the sphere at a rate of $|\boldsymbol{J}|4\pi R^2$ per second, i.e. $\frac{\partial Q(R)}{\partial t} = -|\boldsymbol{J}|4\pi R^2$. On the other hand, the charge $Q(R)$ generates a spherically symmetric electric field for $r > R$:

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\frac{Q(R)}{r^2}\,\boldsymbol{e_r}$$

hence we have

$$\frac{\partial \boldsymbol{E}(\boldsymbol{r})}{\partial t} = -\frac{1}{4\pi\epsilon_0}\frac{1}{r^2}|\boldsymbol{J}|4\pi R^2\,\boldsymbol{e_r} = -\frac{|\boldsymbol{J}|}{\epsilon_0}\frac{R^2}{r^2}\,\boldsymbol{e_r} \qquad \text{for } r \geq R.$$

So on the surface of a sphere of radius R we have

$$\frac{\partial \boldsymbol{E}(R)}{\partial t} = -\frac{|\boldsymbol{J}|}{\epsilon_0}\,\boldsymbol{e_r}$$

hence if we compute

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0\boldsymbol{J} + \epsilon_0\mu_0\frac{\partial \boldsymbol{E}}{\partial t}$$

on the surface of the sphere of radius R, where we have $\boldsymbol{J} = |\boldsymbol{J}|\,\boldsymbol{e_r}$ then we find that

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0|\boldsymbol{J}|\,\boldsymbol{e_r} + \epsilon_0\mu_0(-\frac{|\boldsymbol{J}|}{\epsilon_0}\,\boldsymbol{e_r}) = \boldsymbol{0}$$

and our puzzle is resolved: there is no $\boldsymbol{B}$ field produced by the radial field of current.

We now have the full set of Maxwell's equations:

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0} \qquad \text{Gauss's Law} \tag{6.3}$$

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \qquad \text{Faraday's Law} \tag{6.4}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \tag{6.5}$$

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0\boldsymbol{J} + \epsilon_0\mu_0\frac{\partial \boldsymbol{E}}{\partial t} \qquad \text{Ampère's Law with Maxwell's correction.} \tag{6.6}$$

We notice immediately an interesting feature: a changing electric field produces a magnetic field, and a changing magnetic field in turn produces an electric field. There is an interplay and feedback between the electric and magnetic fields, which was not apparent before Maxwell added the new term to Ampère's law.

6.2 Electromagnetic Waves in Vacuum

Maxwell's equations admit all the solutions to electrostatics and magnetostatics that we have studied so far: because when $\frac{\partial \boldsymbol{E}}{\partial t} = \boldsymbol{0}$ and $\frac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{0}$, Maxwell's equations decouple into the equations for electrostatics and magnetostatics.

In general Maxwell's equations appear challenging to solve. In the face of such a challenge one may try to simplify the problem and solve the simpler problem - this is essentially what we have done in studying electrostatics and magnetostatics. Let us try to solve Maxwell's equations in the absence of charge density (so $\rho = 0$) and in the absence of current density (so $\boldsymbol{J} = \boldsymbol{0}$), which is what we mean talk about studying electrodynamics in a vacuum. Maxwell's equations become

$$\begin{aligned}
\boldsymbol{\nabla} \cdot \boldsymbol{E} &= 0 \\
\boldsymbol{\nabla} \times \boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t} \\
\boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0 \\
\boldsymbol{\nabla} \times \boldsymbol{B} &= \epsilon_0 \mu_0 \frac{\partial \boldsymbol{E}}{\partial t}.
\end{aligned}$$

If we compute

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{E}) = -\boldsymbol{\nabla} \times \frac{\partial \boldsymbol{B}}{\partial t} = -\frac{\partial}{\partial t}(\boldsymbol{\nabla} \times \boldsymbol{B}) = -\epsilon_0 \mu_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}$$

we obtain an equation entirely in terms of the electric field, which we may attempt to solve. Since

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{E}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{E}) - \boldsymbol{\nabla}^2 \boldsymbol{E} = -\boldsymbol{\nabla}^2 \boldsymbol{E}$$

as $\boldsymbol{\nabla} \cdot \boldsymbol{E} = 0$ in the vacuum. Altogether we have

$$\boldsymbol{\nabla}^2 \boldsymbol{E} = \epsilon_0 \mu_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{6.7}$$

Can we find something similar for $\boldsymbol{B}$? Yes we can:

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{B}) = \epsilon_0 \mu_0 \boldsymbol{\nabla} \times \frac{\partial \boldsymbol{E}}{\partial t} = \epsilon_0 \mu_0 \frac{\partial}{\partial t}(\boldsymbol{\nabla} \times \boldsymbol{E}) = -\epsilon_0 \mu_0 \frac{\partial^2 \boldsymbol{B}}{\partial t^2}$$

and since

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{B}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{B}) - \boldsymbol{\nabla}^2 \boldsymbol{B} = -\boldsymbol{\nabla}^2 \boldsymbol{B}$$

because $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$ then we have

$$\boldsymbol{\nabla}^2 \boldsymbol{B} = \epsilon_0 \mu_0 \frac{\partial^2 \boldsymbol{B}}{\partial t^2}. \tag{6.8}$$

These equations have decoupled $\boldsymbol{E}$ and $\boldsymbol{B}$ and are three-dimensional wave equations, one for each component of $\boldsymbol{E}$ and $\boldsymbol{B}$. Let us recall the standard wave equation and its solutions.

6.2.1 Travelling Waves

There are many wave equations, and one tautological way of defining a wave equation is to say that it is an equation which admits travelling waves as a solution. The prototype wave equation in one dimension (with coordinate x) is

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

where c is a constant and $u = u(x,t)$. It admits the solutions $u = f(x-ct) + g(x+ct)$, where f and g are any twice differentiable functions of $x-ct$ and $x+ct$ respectively. Notice that $f(x-ct)$ is a function of shape f which is moving in the positive x direction with speed c, while $g(x+ct)$ is a function of shape g which is moving in the negative x direction with speed c. Both $f(x-ct)$ and $g(x+ct)$ are travelling wave solutions: notice that there is no requirement on them to be sinusoidal functions, any twice differentiable function for f or g will do. Our most important takeaway from these comments is that c, which appears in the wave equation is the speed of the wave.

Of course sinusoidal and cosine functions may be used to make wave solutions, these are called harmonic waves, for example

$$u(x,t) = A\cos(x-ct)$$

is a cosine function of amplitude A travelling at speed c in the positive x direction. If we pause time (set $t=0$ for example) we can read off that the function has wavelength 2π. If we want to consider harmonic wave functions with a different wavelength we must add in a coefficient k, called the wave number which premultiplies x - however since we must consider only functions of $(x-ct)$ any coefficient of x must also be a coefficient of $-ct$ as well, i.e.

$$u(x,t) = A\cos(k(x-ct)) = A\cos(kx-\omega t)$$

where we have defined[6] $\omega = kc$ and now $u(x,t)$ is a travelling wave solution whose wavelength λ is

$$\lambda = \frac{2\pi}{k}.$$

[6]For other wave equations this relation may not be so simple, just how ω depends on k is called a dispersion relation, $\omega(k)$. We may approach any wave equation by substituting a travelling wave solution of the form $A\sin(kx-\omega t)$ then in order to find a travelling wave solution we may find that there will be a non-linear dependence of ω on k. For example the wave equation: $\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} + d\frac{\partial^3 u}{\partial x^3} = 0$ which has a dispersion relation $\omega = ck - dk^3$, so that wavespeed $\frac{\omega}{k} = c - dk^2$ depends on the wavenumber: travelling waves of different wavelength travel at different speeds.

The period $t = \tau$ of the wave is given by $\tau = \frac{2\pi}{\omega}$ and ω is called the angular speed or angular frequency. Everything we have said for our example cosine function applies equally well to a sine function.

It will be helpful to use complex notation to express travelling wave solutions, and to do this we note that

$$u(x,t) = A\cos(k(x-ct)) = A\cos(kx-\omega t) = \mathrm{Re}(Ae^{i(kx-\omega t)})$$

by virtue of Euler's formula. Now by summing travelling waves of different wavelengths and amplitudes one can reconstruct any function[7]:

$$u(x,t) = \mathrm{Re}(\int_{-\infty}^{\infty} A(k)e^{i(kx-\omega t)}\,dk)$$

where by integrating over k we are integrating over the wavelengths for each travelling wave solution. We have made A a function of k so that each travelling wave we add in the integral may have a different amplitude. Since it is cumbersome to work with only the real part of the complex function we will adopt the shorthand

$$\tilde{u}(x,t) = \int_{-\infty}^{\infty} A(k)e^{i(kx-\omega t)}\,dk$$

where it is understood that $u(x,t) = \mathrm{Re}(\tilde{u}(x,t))$.

The main observation is that we can use travelling wave solutions to construct any wave solution by linear superposition, so it will be sufficient for us to focus our interest on travelling wave solutions when solving wave equations.

Finally we may generalise the one-dimensional wave equation to a three-dimensional wave equation:

$$\frac{\partial^2 u}{\partial t^2} = c^2\Big(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\Big) = c^2\boldsymbol{\nabla}^2 u.$$

It is solved by

$$\tilde{u}(\boldsymbol{r},t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(\boldsymbol{k})e^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}\,d^3k$$

[7]To quickly convince yourself of this idea, imagine a sum of exponential functions $\sum_{j=1}^{n} A_j e^{i\omega_j t}$. This corresponds to a sum of n complex numbers of length A_j added together and each rotating at a different rate ω_j. One can view the resulting complex number as the point traced out by a pencil on the circumference of the last wheel in a system of n wheels of radius A_j with wheel j centred on the edge of wheel $j-1$ and so on. The wheels rotate at speeds ω_j and the pencil traces out a complicated path which is $\sum_{j=1}^{n} A_j e^{i\omega_j t}$. It takes some imagination to see this, so why not look at `https://www.youtube.com/watch?v=QVuU2YCwHjw` to get the rough idea. It is the reason why a system of epicycles could be successfully contrived in an attempt to try to explain why the Earth is at the centre of the solar system. Once $n \to \infty$ then the sum becomes an integral and the point drawn out can approximate any reasonable function.

where $\boldsymbol{k} = \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix}$ is called the wave vector, and its components are the wave numbers k_x, k_y and k_z in the x, y and z-directions. We can understand this solution to be a superposition of more fundamental solutions:

$$\tilde{f}(\boldsymbol{r}, t) = A(\boldsymbol{k})e^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}.$$

If $k_y = 0$ and $k_z = 0$ then $\boldsymbol{k} \cdot \boldsymbol{r} = k_x x$ and $\tilde{f}$ is a one-dimensional travelling wave moving in the x-direction, so we see that the wave vector $\boldsymbol{k}$ encodes the wave's direction, as well as its wavelength. By substituting $u = \tilde{f}$ into $\frac{\partial^2 u}{\partial t^2} = c^2\boldsymbol{\nabla}^2 u$ we see that the wave equation is solved by $\tilde{f}$ if $\omega^2 = c^2(k_x^2 + k_y^2 + k_z^2)$, that is if $\omega = c|\boldsymbol{k}|$.

One interesting solution to the wave equation in three dimensions is called a spherical wave which are effectively one-dimensional travelling waves in three-dimensions where the single spatial coordinate is the radial coordinate r. If we rewrite the three-dimensional wave equation in terms of spherical coordinates[8] and consider functions of the form $u(r, t)$ then

$$\begin{aligned}
\frac{\partial^2 u}{\partial t^2} &= c^2\Big(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\Big) \\
&= c^2\Big(\frac{1}{r^2}\frac{\partial}{\partial r}\Big(r^2\frac{\partial u}{\partial r}\Big) + \frac{1}{r^2\cos^2\phi}\frac{\partial^2 u}{\partial\theta^2} + \frac{1}{r^2\cos\phi}\frac{\partial}{\partial\phi}\Big(\cos\phi\frac{\partial u}{\partial\phi}\Big)\Big) \\
&= c^2\frac{1}{r^2}\frac{\partial}{\partial r}\Big(r^2\frac{\partial u}{\partial r}\Big) \\
&= \frac{2c^2}{r}\frac{\partial u}{\partial r} + c^2\frac{\partial^2 u}{\partial r^2}.
\end{aligned}$$

Which has solutions

$$u(r, t) = \frac{1}{r}f(k(r - ct)) + \frac{1}{r}g(k(r + ct))$$

where $\frac{1}{r}f(k(r - ct))$ is a spherical wave travelling radially outwards and $\frac{1}{r}g(k(r + ct))$ is a spherical wave travelling radially inwards.

6.2.2 Plane Wave Solutions of Maxwell's Equations

We have decoupled Maxwell's equations in a vacuum to obtain two second-order differential equations:

$$\begin{aligned}
\frac{\partial^2\boldsymbol{E}}{\partial t^2} &= \frac{1}{\epsilon_0\mu_0}\boldsymbol{\nabla}^2\boldsymbol{E} \\
\frac{\partial^2\boldsymbol{B}}{\partial t^2} &= \frac{1}{\epsilon_0\mu_0}\boldsymbol{\nabla}^2\boldsymbol{B}
\end{aligned}$$

[8] We computed the Laplacian operator in terms of spherical coordinates in section 4.3.

which we now recognise as two three-dimensional wave equations. What is surprising is that both equations have the same wavespeed:

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}.$$

Magnetic waves and electric waves travel at the same speed. Now $\epsilon_0 \approx 8.85418782 \times 10^{-12} \frac{C^2}{Nm^2}$ which is measured using the Coulomb force law, and $\mu_0 = 4\pi \times 10^{-7} \frac{N}{A^2}$ is determined by measuring the strength of the magnetic field due to a constant current using the Biot-Savart law. These quantities were known reasonably well before Maxwell did his work, so he was able to make roughly the same computation that we can and find that[9] $\frac{1}{\sqrt{\epsilon_0 \mu_0}} \approx 2.99792 \times 10^8 ms^{-1}$, which is exactly the speed of light[10]! As Maxwell observed, 'We can scarcely avoid the inference that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena.' Maxwell was referring to the aether, an idea that would later be debunked, but what Maxwell had done through his unification of electricity and magnetism was discover light. Light itself is an electromagnetic wave, and the speed of light c is a constant.

What we have discovered is that in a vacuum each component of the electric and magnetic fields solves its own wave equation, and the waves travel at the speed of light. We still have to make sure the solutions we find are consistent with Maxwell's equations, but we now have a good basis upon which to find these complete solutions. We will start by thinking about a travelling wave solution for a single frequency ω, as the travelling wave solutions can be used to construct the most general solutions of the wave equation. We will further simplify our search by seeking solutions which depend only on one spatial coordinate x and time t and which travel in the positive x direction, i.e. we consider solutions of the form

$$\tilde{\boldsymbol{E}}(x,t) = \tilde{\boldsymbol{E}}_{\boldsymbol{0}} e^{i(k_E x - \omega_E t)} \qquad \text{and} \qquad \tilde{\boldsymbol{B}}(x,t) = \tilde{\boldsymbol{B}}_{\boldsymbol{0}} e^{i(k_B x - \omega_B t)}$$

where $c = \frac{\omega_E}{k_E} = \frac{\omega_B}{k_B}$ and $\tilde{\boldsymbol{E}}_{\boldsymbol{0}}$ and $\tilde{\boldsymbol{B}}_{\boldsymbol{0}}$ are vectors of constant, complex-valued amplitude coefficients. Now consider the equations $\boldsymbol{\nabla} \cdot \tilde{\boldsymbol{E}} = 0$ and $\boldsymbol{\nabla} \cdot \tilde{\boldsymbol{B}} = 0$, if we let $\tilde{\boldsymbol{E}}_{\boldsymbol{0}} = \begin{pmatrix} \tilde{E}_x \\ \tilde{E}_y \\ \tilde{E}_z \end{pmatrix}$ and

[9] For the units, it is helpful to note that 1 Ampère of current is 1 Coulomb of charge moving past a given point per second.

[10] Which had first been measured by Ole Rømer by observing the moons of Jupiter in 1676.

$\tilde{\boldsymbol{B}}_{\mathbf{0}} = \begin{pmatrix} \tilde{B}_x \\ \tilde{B}_y \\ \tilde{B}_z \end{pmatrix}$ then we find

$$\boldsymbol{\nabla} \cdot \tilde{\boldsymbol{E}} = ik_E \tilde{E}_x e^{i(k_E x - \omega_E t)} = 0$$
$$\boldsymbol{\nabla} \cdot \tilde{\boldsymbol{B}} = ik_B \tilde{B}_x e^{i(k_B x - \omega_B t)} = 0$$

hence we deduce that either our solution is trivial so that $k_E = 0$ and $k_B = 0$ (we will see that these will have to vanish together or not at all) or $\tilde{E}_x = 0$ and $\tilde{B}_x = 0$. In the non-trivial case we see that

$$\tilde{\boldsymbol{E}} = \begin{pmatrix} 0 \\ \tilde{E}_y \\ \tilde{E}_z \end{pmatrix} e^{i(k_E x - \omega_E t)} \qquad \text{and} \qquad \tilde{\boldsymbol{B}} = \begin{pmatrix} 0 \\ \tilde{B}_y \\ \tilde{B}_z \end{pmatrix} e^{i(k_B x - \omega_B t)}$$

so that both the electric and magnetic waves have amplitudes which oscillate in directions perpendicular (y and z) to the direction of travel of the wave (x). Such a wave is called a transverse wave and is the type of wave that travels along an oscillating string, its amplitude is in a plane perpendicular to the wave's journey along the string. Alternatively a wave whose amplitudes lie in the direction of the wave, for example think of sound waves or the waves of compression in an oscillating slinky spring or in a line of traffic, are called longitudinal waves. Electromagnetic waves are transverse waves. We must still see if we can solve the remaining pair of Maxwell's equations for $\tilde{\boldsymbol{E}}$ and $\tilde{\boldsymbol{B}}$. We first compute,

$$\boldsymbol{\nabla} \times \tilde{\boldsymbol{E}} = ik_E \begin{pmatrix} 0 \\ -\tilde{E}_z \\ \tilde{E}_y \end{pmatrix} e^{i(k_E x - \omega_E t)}$$

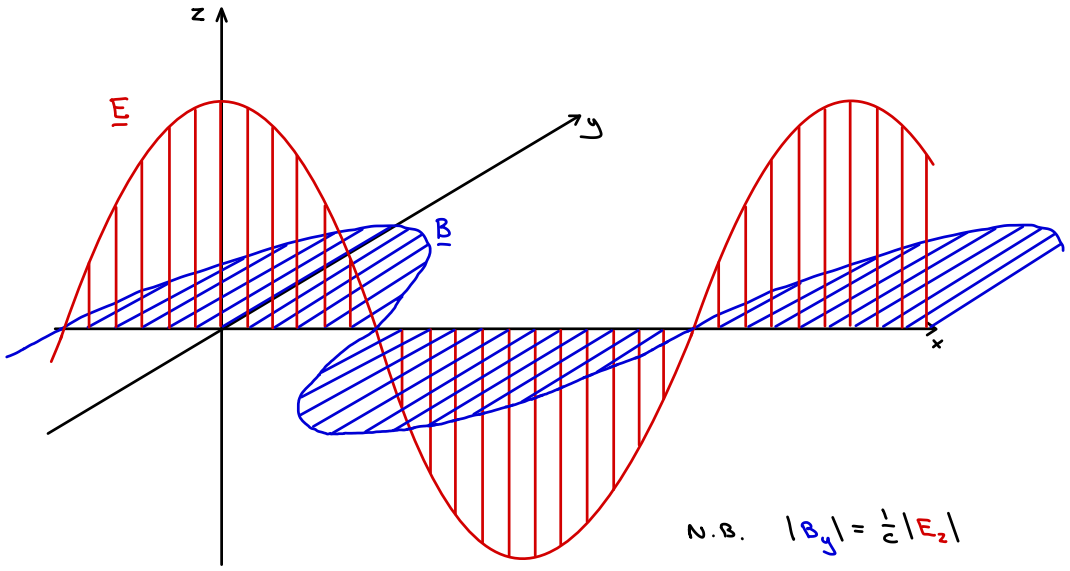
which, to solve Maxwell's equations, must be equal to

$$-\frac{\partial \tilde{B}}{\partial t} = i\omega_B \begin{pmatrix} 0 \\ \tilde{B}_y \\ \tilde{B}_z \end{pmatrix} e^{i(k_B x - \omega_B t)}.$$

For these expressions to be equal each wave must have the same wavelength and frequency, so we must have that $k_E = k_B = k$ and $\omega_E = \omega_B = \omega$. Furthermore we must also have $\tilde{E}_z = -\frac{\omega}{k}\tilde{B}_y = -c\tilde{B}_y$ and $\tilde{E}_y = \frac{\omega}{k}\tilde{B}_z = c\tilde{B}_z$. Altogether we have:

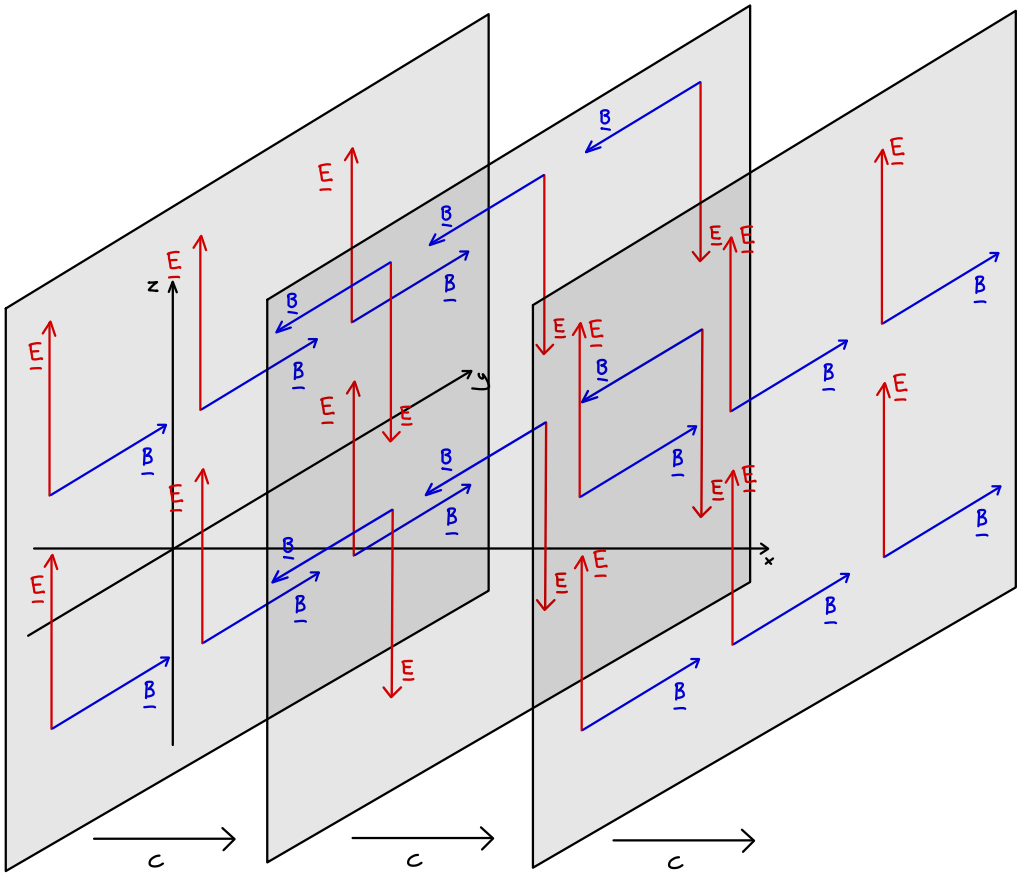
$$\tilde{\boldsymbol{E}} = \begin{pmatrix} 0 \\ \tilde{E}_y \\ \tilde{E}_z \end{pmatrix} e^{i(kx - \omega t)} \qquad \text{and} \qquad \tilde{\boldsymbol{B}} = \frac{1}{c} \begin{pmatrix} 0 \\ -\tilde{E}_z \\ \tilde{E}_y \end{pmatrix} e^{i(kx - \omega t)} \tag{6.9}$$

which one may confirm also satisfies the remaining Maxwell equation in a vacuum: $\boldsymbol{\nabla} \times \tilde{\boldsymbol{B}} = \epsilon_0 \mu_0 \frac{\partial \tilde{\boldsymbol{E}}}{\partial t} = \frac{1}{c^2} \frac{\partial \tilde{\boldsymbol{E}}}{\partial t}$. We notice that $\tilde{\boldsymbol{E}} \cdot \tilde{\boldsymbol{B}} = 0$ so that the electric and magnetic waves are perpendicular to each other, they are in phase and they travel in the x-direction at the speed of light c. The only significant difference beyond their orientation is that the amplitude of the magnetic field is $\frac{1}{c}$ smaller than the amplitude of the electric field. We sketch these ideas below for an electric wave in the xz-plane (i.e. $\tilde{E}_y = 0$, so $\tilde{B}_z = 0$ too), however we suppress the fact that the magnetic wave has an amplitude which is a factor of c smaller than the amplitude of the electric wave, so that the fact that the magnetic and electric waves are in phase and are at right angles to each other may be visible.



We should reflect on what we have drawn above: it is the time-paused profile of an electromagnetic wave along the x-axis, as time flows these waves move at constant speed c, the speed of light, along the x-axis in the positive direction. These are the real parts of $\tilde{\boldsymbol{E}}$ and $\tilde{\boldsymbol{B}}$, remember the complex numbers were just employed for convenience, the real waves are their real parts. We have plotted these waves along the x-axis, but our waves are independent of y and z by design, so, in fact, we would expect these waves to be reproduced around all the lines parallel to the x-axis, we have just sketched the part of the wave about the x-axis. What do the amplitudes tell us? At any point x_0 the amplitude of the electric field drawn above tells us the magnitudes of the electric field in the $\boldsymbol{e_z}$-direction (just for the example above) while the amplitude of the magnetic field at x_0 tells us the magnitude of the magnetic field in the $\boldsymbol{e_y}$-direction (again, just for this particular example).So what we have plotted above tells us the strength and direction of the electric field along all points on the x-axis at a fixed moment in time. Since the solution is independent of y and z it does not change under a translation parallel to the yz-plane. So

if the electric and magnetic fields at a time t_0 and at x_0 are $\boldsymbol{E}(x_0, t_0)$ and $\boldsymbol{B}(x_0, t_0)$ then they take the same values over the entire $x = x_0$ plane at time $t = t_0$. We may plot the positions of the peaks and the troughs of the wave and they would be a series of planes perpendicular to the plane $x = 0$, for this reason this wave solution we have found is called the plane wave, because it gives fields which are constant over a plane at any fixed time.



Plane waves: the peaks of the waves occur in the same plane and move in a direction perpendicular to the plane at speed c.

We have found a plane wave propagating in the x-direction, and we have sketched a special case of this plane wave for which $\tilde{E}_y = 0$. There are two features of this plane wave that we may generalise: the direction of travel, and the plane in which the amplitudes oscillate, which is called the polarisation (in our sketched example we drew a plane wave polarised to oscillate electrically in the z-direction, but we could have given it some component along the y-direction: in three dimensions, electromagnetic waves have two degrees of polarisation). The polarisation of the electromagnetic wave is encoded in the amplitude vectors $\tilde{\boldsymbol{E}}_{\boldsymbol{0}}$ and $\tilde{\boldsymbol{B}}_{\boldsymbol{0}}$. The direction of propagation comes with a wave-number k, to generalise the direction the wave travels in, we

must generalise k to the wave vector $\boldsymbol{k} = \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix}$, so that $\boldsymbol{k} \cdot \boldsymbol{r} = k_x x + k_y y + k_z z$ where k_x, k_y and k_z are the wave numbers in the x, y and z-directions and so encodes the wavelength of the wave in each of these directions. Now it is simple to modify our solution to find a plane wave which propagates in an arbitrary direction $\boldsymbol{k}$ whose polarisation is in the direction $\hat{\boldsymbol{n}}$ (a unit vector):

$$\tilde{\boldsymbol{E}} = \tilde{E}_0 e^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}\hat{\boldsymbol{n}} \qquad \text{and} \qquad \tilde{\boldsymbol{B}} = \frac{1}{c}\tilde{E}_0 e^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}\hat{\boldsymbol{k}} \times \hat{\boldsymbol{n}} \tag{6.10}$$

where $\hat{\boldsymbol{k}}$ is the wave vector $\boldsymbol{k}$ normalised to unit length. Notice that $\tilde{\boldsymbol{B}} = \frac{1}{c}\hat{\boldsymbol{k}} \times \tilde{\boldsymbol{E}}$ so that it is orthogonal to both $\tilde{\boldsymbol{E}}$ and the wave vector $\boldsymbol{k}$ which determines the direction of travel. Since $\boldsymbol{\nabla} \cdot \tilde{\boldsymbol{E}} = 0$ then we find that we are not free to choose $\hat{\boldsymbol{n}}$ and $\boldsymbol{k}$ independently, but we require that $\boldsymbol{k} \cdot \hat{\boldsymbol{n}} = 0$ - this encodes that the electromagnetic waves are transverse waves. The actual electric and magnetic fields of the plane wave are the real parts of these expressions.

The monochromatic plane wave solutions to Maxwell's equations seem particularly special because they are uniform over a plane and they focus on a single value for ω, however they are much more generally useful. Any electromagnetic wave, will be approximated by a plane wave on sufficiently small a scale - so these waves are fundamental in our thinking about more general electromagnetic waves. The monochromatic plane wave solution depends on just a single frequency, or angular speed, ω. Our reasoning for this was that we knew that we could reconstruct a general travelling wave by integrating, or summing, together plane waves for different frequencies. The plane wave above picks out a single frequency, and because the frequency of light corresponds to the interpretation of colours in the human brain, our solution is called a monochromatic plane wave. The connection with visible light was initially built on the observation that $\frac{1}{\sqrt{\epsilon_0 \mu_0}} = c$, the speed of light. However now that we have the plane wave solutions we may use them to test if these electromagnetic waves have the same properties we know of visible light from optics, one can check their properties as they interact with solid media to see if they obey the same laws of reflection, transmission and refraction as visible light. Indeed with a little more work one can derive the fundamental laws of optics from the monochromatic plane wave solution to Maxwell's equation. Experiments show that the correspondence is complete: light is an electromagnetic wave. What is more there are frequencies of the electromagnetic wave which go beyond the corresponding frequencies for coloured light. At the low frequency end of the electromagnetic spectrum one finds the kinds of electromagnetic waves which are used for transmitting radio and television signals, while at the high frequency end one discovers ultraviolet light, x-rays and gamma rays which may be distinguished among the continuous spectrum of frequencies. All of this came from Maxwell's

unification of the theories of the electric and magnetic fields, it was an immense leap in human understanding, one which we are still using and developing to this day.

6.3 Electromagnetic Energy

We have thought about the work one needs to do to overcome the Coulomb force and construct a system of charges in electrostatics, this is the potential energy put into and so contained in the system of charges in electrostatics

$$W_E = \frac{1}{2}\sum_{i=1}^{n} q_i \Phi(\boldsymbol{r_i'}).$$

We may develop this, still in the setting of electrostatics, into an integral for a volume charge density ρ for a system contained in a volume V with electric potential Φ:

$$W_E = \frac{1}{2}\int_V \rho\Phi \, dv$$

and we may go further and rewrite this in terms of the electric field $\boldsymbol{E}$. We will first constrain our thinking to electrostatics. Now we may use $\boldsymbol{\nabla}\cdot\boldsymbol{E} = \frac{\rho}{\epsilon_0}$ to first eliminate ρ:

$$W_E = \frac{\epsilon_0}{2}\int_V (\boldsymbol{\nabla}\cdot\boldsymbol{E})\Phi \, dv$$

and then integration by parts gives us

$$\begin{aligned} W_E &= -\frac{\epsilon_0}{2}\int_V \boldsymbol{E}\cdot\boldsymbol{\nabla}\Phi \, dv + \frac{\epsilon_0}{2}\int_V \boldsymbol{\nabla}\cdot(\boldsymbol{E}\Phi)\, dv \\ &= -\frac{\epsilon_0}{2}\int_V \boldsymbol{E}\cdot\boldsymbol{\nabla}\Phi \, dv + \frac{\epsilon_0}{2}\oint_S (\boldsymbol{E}\Phi)\cdot d\boldsymbol{a} \\ &= \frac{\epsilon_0}{2}\int_V \boldsymbol{E}^2 \, dv + \frac{\epsilon_0}{2}\oint_S (\boldsymbol{E}\Phi)\cdot d\boldsymbol{a} \end{aligned}$$

What happens when we extend V to cover all space? Since we defined V to be a volume containing our charge density ρ then we do not expect any change in W. However since the electric field does not vanish, in our expression above the first integral, which is positive-definite grows, and so the second term must decrease (to keep W constant). Now far from the charge density (out towards infinity) then in the surface integral $\boldsymbol{E} \propto \frac{1}{r^2}$, $\Phi \propto \frac{1}{r}$ and (in spherical coordinates) $d\boldsymbol{a} \propto r^2$ so the second integral overall is proportional to $\frac{1}{r}$. If we integrate over all space then $r \to \infty$, so the second terms decreases all the way to zero if we take V to be $\mathbb{R}^3$, all of space then the electrostatic energy in the field $\boldsymbol{E}$ is

$$W_E = \frac{\epsilon_0}{2}\int_{\mathbb{R}^3} \boldsymbol{E}^2 \, dv. \tag{6.11}$$

We will now find a similar expression for the magnetostatic energy contained in the field $\boldsymbol{B}$. What is the corresponding situation in magnetostatics to assembling a configuration of point charges? It is assembling a system of steady currents. But, wait, we have already proved in example 5.3 that the total force on any loop of current due to a static magnetic field is zero, so what is there to do work against? There is another impediment at work, recall that when we discussed the electromotive force we realised that changing a steady current introduces a changing magnetic field which induces a current in the wire that opposes the change in the initial current. These electromotive forces that a changing current induces are what we must do work against to construct a loop of current. What can we say about them? A current I flows in a wire, and produces a magnetic field which, according to the Biot-Savart law is perpendicular to the current and also proportional to the current. In turn the magnetic field has flux through the area bounded by the loop of wire so $\Phi_B = \int_S \boldsymbol{B} \cdot d\boldsymbol{a}$ where S is the surface bounded by the loop of wire carrying the current I, as $\boldsymbol{B}$ is proportional to the changing current then then the (back) electromotive force induced in a wire by a changing current is

$$\mathcal{E} = -\frac{d\Phi_B}{dt} \propto -\frac{dI}{dt}$$

the constant of proportionality is called the self-inductance of the current loop, and we denote it by L, hence

$$\Phi_B = LI.$$

The self-inductance L depends on the shape of the loop of wire, it is a geometrical quantity. Now to set up a current in a loop of wire we must overcome the induced (back) electromotive force, since we are doing work against the electromotive force we must do work per unit charge $-\mathcal{E}$ and the current tells us how much charge we are moving, in time dt we move $I\,dt$ charge, i.e. $dW = -\mathcal{E} I\,dt$ or,

$$\frac{dW_B}{dt} = -\mathcal{E}I = I\frac{d\Phi_B}{dt} = LI\frac{dI}{dt}.$$

Integrating we find that

$$W_B = \frac{1}{2}LI^2$$

this is the energy stored in a current in magnetostatics. Now we may rewrite this expression and generalise it to volume current densities. Note that

$$LI = \Phi_B = \int_S \boldsymbol{B} \cdot d\boldsymbol{a} = \int_S (\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot d\boldsymbol{a} = \oint_P \boldsymbol{A} \cdot d\boldsymbol{l}$$

where P is the path of the wire carrying the current, and so we have

$$W_B = \frac{1}{2} I \oint_P \boldsymbol{A} \cdot d\boldsymbol{l} = \frac{1}{2} \oint_P (\boldsymbol{A} \cdot \boldsymbol{I})\, dl$$

where we have used $\boldsymbol{I} = I d\boldsymbol{l}$ along an element of the wire. From here we may generalise to consider the energy contained in a surface current or a volume current, e.g.

$$W_B = \frac{1}{2}\int_V (\boldsymbol{A}\cdot\boldsymbol{J})\, dv.$$

Now just as we did for the energy in a static electric field, we may use the equations governing magnetostatics to rewrite the energy above in terms of the magnetic field. Ampére's law $\boldsymbol{\nabla}\times\boldsymbol{B} = \mu_0\boldsymbol{J}$ allows us to eliminate $\boldsymbol{J}$:

$$W_B = \frac{1}{2\mu_0}\int_V (\boldsymbol{A}\cdot(\boldsymbol{\nabla}\times\boldsymbol{B}))\, dv.$$

Now because $\boldsymbol{\nabla}$ is an operator we cannot simply use the scalar triple product identities to rewrite this in terms of just $\boldsymbol{B}$, we must be a little more cautious only to obtain the same result but for correct reasons. We must use integration by parts to obtain

$$\begin{aligned} W_B &= \frac{1}{2\mu_0}\int_V \boldsymbol{B}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A})\, dv - \frac{1}{2\mu_0}\int_V \boldsymbol{\nabla}\cdot(\boldsymbol{A}\times\boldsymbol{B})\, dv \\ &= \frac{1}{2\mu_0}\int_V \boldsymbol{B}^2\, dv - \frac{1}{2\mu_0}\int_S (\boldsymbol{A}\times\boldsymbol{B})\cdot d\boldsymbol{a}. \end{aligned}$$

If we allow V to grow to cover all space then the terms in the surface integral are in spherical coordinates, altogether, proportional to $\frac{1}{r}$ and so the surface integral goes to zero as $r\to\infty$ when we take $V\to\mathbb{R}^3$. Hence we find that over all space the energy in a static magnetic field $\boldsymbol{B}$ is

$$W_B = \frac{1}{2\mu_0}\int_{\mathbb{R}^3} \boldsymbol{B}^2\, dv. \tag{6.12}$$

This is pleasingly similar to the expression for the energy in the static electric field.

The total energy in a configuration static electric and static magnetic fields is $W_E + W_B$, what changes when we consider electrodynamics? Even though some of the expressions we used to derive the expressions for W_E and W_B are no longer valid in electrodynamics - it is only the time-dependent terms whose role has been omitted. If we pause a dynamic electromagnetic system then the energy in the system is given by $W_E + W_B$. When we consider the dynamical system, we may hypothesis two mechanisms through which the electromagnetic fields can lose energy. First, the fields may do mechanical work on charge: the Lorentz force law tells us the force that $\boldsymbol{E}$ and $\boldsymbol{B}$ bring to bear on a test charge Q e.g. the electric and magnetic fields may cause currents to move around a wire, but if the wire is not a perfect conductor then it will offer resistance to the motion of the charges and to keep the current steady will require work from the electromagnetic fields and energy will be lost from the fields as heat in the wire

(meaning kinetic energy in the particles in the wire). This is an example of mechanical energy loss in an electromagnetic system. A second way we can imagine energy being removed from the electromagnetic fields is if electromagnetic waves carry energy away in their oscillations, we may call this radiation energy loss.

Let us first consider the mechanical energy loss by computing the work done by $\boldsymbol{E}$ and $\boldsymbol{B}$ on a test charge Q. The Lorentz force law tells us the force $\boldsymbol{F} = Q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$ done on a charge Q moving with velocity $\boldsymbol{v}$. In a time dt this charge Q will move $d\boldsymbol{l} = \boldsymbol{v}dt$, so that $\boldsymbol{F} \cdot d\boldsymbol{l} = Q\boldsymbol{E} \cdot \boldsymbol{v}dt$. For a charge distribution ρ we find that

$$\frac{dW}{dt} = \int_V \rho \boldsymbol{E} \cdot \boldsymbol{v}\, dv = \int_V \boldsymbol{E} \cdot (\rho \boldsymbol{v})\, dv = \int_V (\boldsymbol{E} \cdot \boldsymbol{J})\, dv$$

this is the rate of change of the work done, or the rate of energy loss of the electromagnetic fields. We can use Maxwell's modification of Ampère's law (equation (6.6)) to eliminate the current density in favour of $\boldsymbol{B}$ and $\boldsymbol{E}$: substituting $\boldsymbol{J} = \frac{1}{\mu_0}(\boldsymbol{\nabla} \times \boldsymbol{B}) - \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}$ gives

$$\begin{aligned}
\frac{dW}{dt} &= \frac{1}{\mu_0} \int_V \boldsymbol{E} \cdot (\boldsymbol{\nabla} \times \boldsymbol{B})\, dv - \epsilon_0 \int_V \boldsymbol{E} \cdot \frac{\partial \boldsymbol{E}}{\partial t}\, dv \\
&= \frac{1}{\mu_0} \int_V \boldsymbol{B} \cdot (\boldsymbol{\nabla} \times \boldsymbol{E})\, dv - \frac{1}{\mu_0} \int_V \boldsymbol{\nabla} \cdot (\boldsymbol{E} \times \boldsymbol{B})\, dv - \epsilon_0 \int_V \boldsymbol{E} \cdot \frac{\partial \boldsymbol{E}}{\partial t}\, dv \\
&= -\frac{1}{\mu_0} \int_V \boldsymbol{B} \cdot \frac{\partial \boldsymbol{B}}{\partial t}\, dv - \frac{1}{\mu_0} \int_V \boldsymbol{\nabla} \cdot (\boldsymbol{E} \times \boldsymbol{B})\, dv - \epsilon_0 \int_V \boldsymbol{E} \cdot \frac{\partial \boldsymbol{E}}{\partial t}\, dv \\
&= -\frac{d}{dt}\Big(\frac{1}{2} \int_V \big(\epsilon_0 \boldsymbol{E}^2 + \frac{1}{\mu_0} \boldsymbol{B}^2\big)\, dv\Big) - \frac{1}{\mu_0} \oint_S (\boldsymbol{E} \times \boldsymbol{B}) \cdot d\boldsymbol{a}.
\end{aligned}$$

Rearranging we have,

$$\frac{d}{dt}\Big(\frac{1}{2} \int_V \big(\epsilon_0 \boldsymbol{E}^2 + \frac{1}{\mu_0} \boldsymbol{B}^2\big)\, dv\Big) = -\frac{dW}{dt} - \frac{1}{\mu_0} \oint_S (\boldsymbol{E} \times \boldsymbol{B}) \cdot d\boldsymbol{a}.$$

The left-hand-side is the rate of change of energy in the electromagnetic fields in the volume V. Now we see the two distinct sources of loss of energy in the electromagnetic fields: the first terms $-\frac{dW}{dt} = -\int_V (\boldsymbol{E} \cdot \boldsymbol{J})\, dv$ tells us about the mechanical energy loss in the volume V, and the second integral tells us about the loss of energy through the boundary S, which must be due to electromagnetic waves radiating energy away. The mathematical statement above concerning the loss of energy in an electromagnetic system is called Poynting's theorem and the term in the surface integral:

$$\boldsymbol{S} = \frac{1}{\mu_0}(\boldsymbol{E} \times \boldsymbol{B})$$

is called the Poynting vector, and since $\frac{1}{\mu_0} \oint_S (\boldsymbol{E} \times \boldsymbol{B}) \cdot d\boldsymbol{a} = \oint_S \boldsymbol{S} \cdot d\boldsymbol{a}$ represents the loss of electromagnetic energy through a surface S, then $\boldsymbol{S}$ is the energy flux density. Let us simplify

the statement of Poynting's theorem by writing $U = \frac{1}{2}(\epsilon_0 \boldsymbol{E}^2 + \frac{1}{\mu_0}\boldsymbol{B}^2)$ for the energy density in V then we have

$$\frac{d}{dt}\int_V U\, dv = -\frac{dW}{dt} - \oint_S \boldsymbol{S}\cdot d\boldsymbol{a}$$
$$\int_V \frac{\partial U}{\partial t}\, dv = -\frac{dW}{dt} - \int_V (\boldsymbol{\nabla}\cdot\boldsymbol{S})\, dv.$$

Now if there is no mechanical work done on the charges in V, the $\frac{dW}{dt} = 0$ and we are left with

$$\frac{\partial U}{\partial t} = -\boldsymbol{\nabla}\cdot\boldsymbol{S}$$

which is a continuity equation for the electromagnetic energy (when $\frac{dW}{dt} = 0$). It tells us that when energy is not created nor destroyed within a volume ($\frac{dW}{dt} = 0$), then its rate of decrease is given only by the energy flux over the surface of the volume. Of course the assumption that $\frac{dW}{dt} = 0$ is a theoretical one, and in a realistic situation electromagnetic energy can be converted to other forms of energy through the fields doing work on the charges.

6.4 Potentials

In electrostatics, we were able to make use of $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$, only because we had $\boldsymbol{\nabla}\times\boldsymbol{E} = \boldsymbol{0}$. This is no longer the case in electrodynamics. On the other hand we do still have $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$ and so we may again write $\boldsymbol{B} = \boldsymbol{\nabla}\times\boldsymbol{A}$ and try to use this to reformulate Maxwell's equations in terms of potentials. However now that we are considering dynamic fields we will need to assume that $\boldsymbol{A}$ is a functions of time as well as the spatial coordinates. We will first develop Faraday's law given in equation (6.4)

$$\boldsymbol{\nabla}\times\boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} = -\frac{\partial}{\partial t}(\boldsymbol{\nabla}\times\boldsymbol{A}) = -\boldsymbol{\nabla}\times\frac{\partial \boldsymbol{A}}{\partial t}$$

so that

$$\boldsymbol{\nabla}\times\left(\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t}\right) = \boldsymbol{0}.$$

This gives us a way to re-introduce the electric scalar potential which will now be defined by

$$\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} = -\boldsymbol{\nabla}\Phi$$

and we see that for a static magnetic field this reduces to the equation $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$ we have used to develop electrostatics. The problem of solving Maxwell's equations is now reduced to solving Gauss's law in equation (6.3) and Maxwell's modification of Ampère's law in equation

(6.6). The other two of Maxwell's equations are automatically solved by the conventions used to define the potentials. From Gauss's law we obtain

$$\boldsymbol{\nabla}\cdot\boldsymbol{E} = -\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\Phi + \frac{\partial \boldsymbol{A}}{\partial t}) = -\boldsymbol{\nabla}^2\Phi - \frac{\partial}{\partial t}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) = \frac{\rho}{\epsilon_0} \tag{6.13}$$

and from Ampère's law with Maxwell's term:

$$\boldsymbol{\nabla}\times\boldsymbol{B} = \boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A}) = \mu_0\boldsymbol{J} - \epsilon_0\mu_0\frac{\partial}{\partial t}(\boldsymbol{\nabla}\Phi + \frac{\partial \boldsymbol{A}}{\partial t}) = \mu_0\boldsymbol{J} - \epsilon_0\mu_0\boldsymbol{\nabla}\Big(\frac{\partial\Phi}{\partial t}\Big) - \epsilon_0\mu_0\frac{\partial^2\boldsymbol{A}}{\partial t^2}.$$

Since $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \boldsymbol{\nabla}^2\boldsymbol{A}$ then we may rearrange the equation above to obtain

$$\boldsymbol{\nabla}\Big(\boldsymbol{\nabla}\cdot\boldsymbol{A} + \epsilon_0\mu_0\frac{\partial\Phi}{\partial t}\Big) - \boldsymbol{\nabla}^2\boldsymbol{A} + \epsilon_0\mu_0\frac{\partial^2\boldsymbol{A}}{\partial t^2} = \mu_0\boldsymbol{J}. \tag{6.14}$$

Unfortunately equations (6.13) and (6.14) do not appear simple to solve, and they certainly do not have the form of Poisson's equation which we have some practise at solving. We still have a trick at our disposal, namely the gauge invariance of $\boldsymbol{A}$, that we can add a term $\boldsymbol{\nabla}\Lambda$ to $\boldsymbol{A}$ and still obtain the same magnetic field. However we note that changing $\boldsymbol{A}\to\boldsymbol{A}+\boldsymbol{\nabla}\Lambda$ now has an effect on the scalar potential:

$$-\boldsymbol{\nabla}\phi = \boldsymbol{E} + \frac{\partial\boldsymbol{A}}{\partial t} \longrightarrow \boldsymbol{E} + \frac{\partial\boldsymbol{A}}{\partial t} + \frac{\partial}{\partial t}(\boldsymbol{\nabla}\Lambda) = -\boldsymbol{\nabla}(\Phi - \frac{\partial\Lambda}{\partial t}).$$

If we wish to leave $\boldsymbol{E} + \frac{\partial\boldsymbol{A}}{\partial t}$ unchanged under $\boldsymbol{A}\to\boldsymbol{A}+\boldsymbol{\nabla}\Lambda$, then we require the electric potential to also transform at the same time so that $\Phi\to\Phi - \frac{\partial\Lambda}{\partial t}$. The gauge transformations are intertwined.

In magnetostatics, we made use of the Coulomb gauge for which we choose Λ so that $\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0$, in this case equation (6.13) becomes

$$-\boldsymbol{\nabla}^2\Phi = \frac{\rho}{\epsilon_0}$$

which we can solve (in principle), while equation (6.14) becomes

$$\epsilon_0\mu_0\boldsymbol{\nabla}\Big(\frac{\partial\Phi}{\partial t}\Big) - \boldsymbol{\nabla}^2\boldsymbol{A} + \epsilon_0\mu_0\frac{\partial^2\boldsymbol{A}}{\partial t^2} = \mu_0\boldsymbol{J}$$

which is not so simple to solve.

Fortunately there is another choice of gauge, which is called the Lorenz gauge[11], for which we pick Λ so that

$$\boldsymbol{\nabla}\cdot\boldsymbol{A} = -\epsilon_0\mu_0\frac{\partial\Phi}{\partial t}$$

[11]This is not a typo. A typo would be to give this the name Lorentz gauge, after Hendrik Lorentz, rather than to credit it to Ludwig Lorenz. It just so happens that the Lorenz gauge is Lorentz invariant, a sentence we will be better able to understand after we meet the Lorentz transformations.

in order to simplify equation (6.14) to read

$$\boldsymbol{\nabla}^2\boldsymbol{A} - \epsilon_0\mu_0\frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\mu_0\boldsymbol{J}.$$

While equation (6.13) becomes

$$\boldsymbol{\nabla}^2\Phi - \epsilon_0\mu_0(\frac{\partial^2\Phi}{\partial t^2}) = -\frac{\rho}{\epsilon_0}$$

and the two Maxwell equations for the potentials have not only decoupled but also look rather similar. In fact if we define the operator, called the d'Alembertian, $\Box = \boldsymbol{\nabla}^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}$ (recalling that $\epsilon_0\mu_0 = \frac{1}{c^2}$, where c is the speed of light) then Maxwell's equations in Lorenz gauge reduce to

$$\Box\Phi = -\frac{\rho}{\epsilon_0} \tag{6.15}$$

$$\Box\boldsymbol{A} = -\mu_0\boldsymbol{J}. \tag{6.16}$$

The d'Alembertian is the natural extension of the Laplacian to a four dimensional space-time operator. The simplicity of the equations, especially the fact that the four equations can be written in this way, is an indication that in electromagnetism the nature of space and time is different to the absolute space of Newton where time parameterises the dynamics.

6.5 The Retarded and the Advanced Potentials

Equations (6.15) and (6.16) are similar to Poisson's equation, indeed for the time-independent case these equations become Poisson's equation for the scalar potential and its equivalent for the vector potential. Even with the time-dependent terms remaining we can reorder the equations so that the Laplacian appears on the left-hand-side of each, and in that case $\frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2}$ appears as an additional source of charge density, and $\frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2}$ is interpreted as an additional source of current density. However this does not admit a straightforward solution via the Green's function for the Laplacian. In addition we need to address another question: both ρ and $\boldsymbol{J}$ may be time-dependent, so for which moment in time should we measure the charge distribution and current density? The answer is deceptively simple, we know that electromagnetic signals travel at the speed of light. What will matter is not where the charges and currents are now, but where they were when an electromagnetic signal was sent some time before which only reaches the test particle at the present time. Since the distance between a test charge Q at $\boldsymbol{r}$ and the source charge or current at $\boldsymbol{r'}$ is the displacement $s = |\boldsymbol{s}| = |\boldsymbol{r} - \boldsymbol{r'}|$ then a signal from the source to

the test charge will take a time $\frac{s}{c}$ to reach $\boldsymbol{r}$ from $\boldsymbol{r'}$ so we must measure ρ and $\boldsymbol{J}$ at an earlier time t_R which is called the retarded time (meaning the delayed time) and given by

$$t_R = t - \frac{s}{c}$$

where t is the present time for the test charge. It will turn out that

$$\Phi(\boldsymbol{r}, t) = \frac{1}{4\pi\epsilon_0} \int_V \frac{\rho(\boldsymbol{r'}, t_R)}{s}\, dv' \tag{6.17}$$

$$\boldsymbol{A}(\boldsymbol{r}, t) = \frac{\mu_0}{4\pi} \int_V \frac{\boldsymbol{J}(\boldsymbol{r'}, t_R)}{s}\, dv' \tag{6.18}$$

and we note that the more distant $\boldsymbol{r'}$ is, the longer ago we must evaluate ρ and $\boldsymbol{J}$: t_R varies across a distribution of charge or current. The equations above are called the retarded potentials for the distribution. Let us show how these retarded potentials solve equations (6.15) and (6.16) and satisfy the Lorenz gauge condition. Now we recall that $\boldsymbol{\nabla}^2(\frac{1}{s}) = -4\pi\delta^3(\boldsymbol{s})$ so we will aim to separate out these terms. We will focus on equation (6.15) for the electric potential and recalling that both s and t_R are functions of $\boldsymbol{r}$ and that $\rho = \rho(\boldsymbol{r'}, t_R)$ we compute

$$\begin{aligned}
\boldsymbol{\nabla}^2\Phi(\boldsymbol{r}, t) &= \frac{1}{4\pi\epsilon_0}\boldsymbol{\nabla} \cdot \Big(\int_V \Big(\boldsymbol{\nabla}(\frac{1}{s})\rho + \frac{1}{s}(\boldsymbol{\nabla}\rho) \Big)\, dv' \Big) \\
&= \frac{1}{4\pi\epsilon_0} \int_V \Big(\boldsymbol{\nabla}^2(\frac{1}{s})\rho + \boldsymbol{\nabla}(\frac{1}{s}) \cdot (\boldsymbol{\nabla}\rho) + \boldsymbol{\nabla}(\frac{1}{s}) \cdot (\boldsymbol{\nabla}\rho) + \frac{1}{s}\boldsymbol{\nabla}^2\rho \Big)\, dv' \\
&= \frac{1}{4\pi\epsilon_0} \int_V \Big(-4\pi\delta^3(\boldsymbol{s})\rho + 2\boldsymbol{\nabla}(\frac{1}{s}) \cdot (\boldsymbol{\nabla}\rho) + \frac{1}{s}\boldsymbol{\nabla}^2\rho \Big)\, dv'
\end{aligned}$$

where we have used $\boldsymbol{\nabla} \cdot (f\boldsymbol{v}) = (\boldsymbol{\nabla} f) \cdot \boldsymbol{v} + f(\boldsymbol{\nabla} \cdot \boldsymbol{v})$ for a scalar function f and a vector function $\boldsymbol{v}$. Now we know that $\boldsymbol{\nabla}(\frac{1}{s}) = -\frac{1}{s^3}\boldsymbol{s}$, but we must compute $\boldsymbol{\nabla}\rho$ and $\boldsymbol{\nabla}^2\rho$. We have

$$\boldsymbol{\nabla}\rho(\boldsymbol{r'}, t_R) = \frac{\partial\rho}{\partial t_R}\boldsymbol{\nabla}t_R = \frac{\partial\rho}{\partial t_R}\boldsymbol{\nabla}(t - \frac{s}{c}) = -\frac{\partial\rho}{\partial t_R}\frac{1}{cs}\boldsymbol{s} = -\frac{\dot{\rho}}{cs}\boldsymbol{s}$$

where we have used that $\frac{\partial}{\partial x}(t_R) = -\frac{x-x'}{sc}$, $\frac{\partial}{\partial y}(t_R) = -\frac{y-y'}{sc}$ and $\frac{\partial}{\partial z}(t_R) = -\frac{z-z'}{sc}$. In the last equality we have written $\dot{\rho} = \frac{\partial\rho}{\partial t}$, to see this is a valid replacement we have used the definition

of the retarded time $t_R = t - \frac{s}{c}$ to deduce that $\frac{\partial}{\partial t_R} = \frac{\partial}{\partial t}$. We also need

$$\begin{aligned}
\boldsymbol{\nabla}^2(\rho(\boldsymbol{r'}, t_R)) &= \boldsymbol{\nabla} \cdot (-\frac{\dot{\rho}}{cs}\boldsymbol{s}) \\
&= \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \cdot \begin{pmatrix} -\frac{x-x'}{cs}\dot{\rho} \\ -\frac{y-y'}{cs}\dot{\rho} \\ -\frac{z-z'}{cs}\dot{\rho} \end{pmatrix} \\
&= -\frac{1}{cs}\dot{\rho} + \frac{(x-x')^2}{cs^3}\dot{\rho} + \frac{(x-x')^2}{(cs)^2}\ddot{\rho} \\
&\quad - \frac{1}{cs}\dot{\rho} + \frac{(y-y')^2}{cs^3}\dot{\rho} + \frac{(y-y')^2}{(cs)^2}\ddot{\rho} \\
&\quad - \frac{1}{cs}\dot{\rho} + \frac{(z-z')^2}{cs^3}\dot{\rho} + \frac{(z-z')^2}{(cs)^2}\ddot{\rho} \\
&= -\frac{2}{cs}\dot{\rho} + \frac{1}{c^2}\ddot{\rho}.
\end{aligned}$$

Finally we note that

$$\int_V \delta^3(\boldsymbol{s})\rho(\boldsymbol{r'}, t_R)\, dv' = \int_V \delta^3(\boldsymbol{s})\rho(\boldsymbol{r'}, t - \frac{s}{c})\, dv' = \rho(\boldsymbol{r}, t).$$

Putting this altogether we find,

$$\begin{aligned}
\boldsymbol{\nabla}^2\Phi(\boldsymbol{r}, t) &= \frac{1}{4\pi\epsilon_0}\int_V \Big(-4\pi\delta^3(\boldsymbol{s})\rho + 2\boldsymbol{\nabla}(\frac{1}{s}) \cdot (\boldsymbol{\nabla}\rho) + \frac{1}{s}\boldsymbol{\nabla}^2\rho \Big)\, dv' \\
&= -\frac{\rho(\boldsymbol{r}, t)}{\epsilon_0} + \frac{1}{4\pi\epsilon_0}\int_V \Big(2(-\frac{1}{s^3}\boldsymbol{s}) \cdot (-\frac{\dot{\rho}}{cs}\boldsymbol{s}) - \frac{2}{cs^2}\dot{\rho} + \frac{1}{sc^2}\ddot{\rho} \Big)\, dv' \\
&= -\frac{\rho(\boldsymbol{r}, t)}{\epsilon_0} + \frac{1}{c^2}\frac{1}{4\pi\epsilon_0}\int_V \Big(\frac{\ddot{\rho}}{s} \Big)\, dv' \\
&= -\frac{\rho(\boldsymbol{r}, t)}{\epsilon_0} + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\Phi(\boldsymbol{r}, t).
\end{aligned}$$

Rearranging we find $\Box\Phi = -\frac{\rho}{\epsilon_0}$. The computation to show that the retarded magnetic potential solves each component of equation (6.16) is identical up to factors. We may check that we are working in the Lorenz gauge. To do this we will need swap $\boldsymbol{\nabla}$ for $\boldsymbol{\nabla'}$ in places, where $\boldsymbol{\nabla'} = \begin{pmatrix} \frac{\partial}{\partial x'} \\ \frac{\partial}{\partial y'} \\ \frac{\partial}{\partial z'} \end{pmatrix}$, in particular we note that $\boldsymbol{\nabla}(s) = \frac{\boldsymbol{s}}{s}$, $\boldsymbol{\nabla'}(s) = -\frac{\boldsymbol{s}}{s}$,

$$\begin{aligned}
\boldsymbol{\nabla} \cdot \Big(\frac{\boldsymbol{J}}{s}\Big) &= \boldsymbol{\nabla}(\frac{1}{s}) \cdot \boldsymbol{J} + \frac{1}{s}\boldsymbol{\nabla} \cdot \boldsymbol{J}, \qquad \text{and} \\
\boldsymbol{\nabla'} \cdot \Big(\frac{\boldsymbol{J}}{s}\Big) &= \boldsymbol{\nabla'}(\frac{1}{s}) \cdot \boldsymbol{J} + \frac{1}{s}\boldsymbol{\nabla'} \cdot \boldsymbol{J}.
\end{aligned}$$

Further we note that

$$\boldsymbol{\nabla}\cdot\boldsymbol{J}(\boldsymbol{r'},t_R) = -\frac{1}{c}\dot{\boldsymbol{J}}\cdot\boldsymbol{\nabla}(s)$$
$$\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{r'},t_R) = -\dot{\rho} - \frac{1}{c}\dot{\boldsymbol{J}}\cdot\boldsymbol{\nabla}'(s)$$

where we note that the dependence on x, y and z arises only through the appearance of s in the retarded time, and we have used the continuity equation for the charge i.e. for fixed t_R we have $\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{r'}) = -\dot{\rho}$. Now we may compute $\boldsymbol{\nabla}\cdot\boldsymbol{A}$ to check that we are working in the Lorenz gauge:

$$\begin{aligned}
\boldsymbol{\nabla}\cdot\boldsymbol{A} &= \boldsymbol{\nabla}\cdot\Big(\frac{\mu_0}{4\pi}\int_V \frac{\boldsymbol{J}}{s}\,dv'\Big)\\
&= \frac{\mu_0}{4\pi}\int_V \boldsymbol{\nabla}\cdot\Big(\frac{\boldsymbol{J}}{s}\Big)\,dv'\\
&= \frac{\mu_0}{4\pi}\int_V \Big(\boldsymbol{\nabla}(\frac{1}{s})\cdot\boldsymbol{J} + \frac{1}{s}\boldsymbol{\nabla}\cdot\boldsymbol{J} + \boldsymbol{\nabla}'(\frac{1}{s})\cdot\boldsymbol{J} + \frac{1}{s}\boldsymbol{\nabla}'\cdot\boldsymbol{J} - \boldsymbol{\nabla}'\cdot\Big(\frac{\boldsymbol{J}}{s}\Big)\Big)\,dv'\\
&= \frac{\mu_0}{4\pi}\int_V \Big(\frac{1}{s}\boldsymbol{\nabla}\cdot\boldsymbol{J} + \frac{1}{s}\boldsymbol{\nabla}'\cdot\boldsymbol{J} - \boldsymbol{\nabla}'\cdot\Big(\frac{\boldsymbol{J}}{s}\Big)\Big)\,dv'\\
&= \frac{\mu_0}{4\pi}\int_V \Big(-\frac{1}{sc}\dot{\boldsymbol{J}}\cdot\boldsymbol{\nabla}(s) - \frac{1}{s}\dot{\rho} - \frac{1}{sc}\dot{\boldsymbol{J}}\cdot\boldsymbol{\nabla}'(s) - \boldsymbol{\nabla}'\cdot\Big(\frac{\boldsymbol{J}}{s}\Big)\Big)\,dv'\\
&= \frac{\mu_0}{4\pi}\int_V \Big(-\frac{1}{s}\dot{\rho} - \boldsymbol{\nabla}'\cdot\Big(\frac{\boldsymbol{J}}{s}\Big)\Big)\,dv'\\
&= -\epsilon_0\mu_0\frac{\partial\Phi}{\partial t} - \int_V \Big(\boldsymbol{\nabla}'\cdot\Big(\frac{\boldsymbol{J}}{s}\Big)\Big)\,dv'
\end{aligned}$$

which gives the Lorenz gauge whenever $\int_V \Big(\boldsymbol{\nabla}'\cdot\Big(\frac{\boldsymbol{J}}{s}\Big)\Big)\,dv' = \oint_S \frac{\boldsymbol{J}}{s}\cdot d\boldsymbol{a} = 0$ i.e. if the current density is zero over the boundary at infinity.

Therefore the retarded potentials given in equations (6.17) and (6.18) do satisfy Maxwell's equations, and for a given charge density ρ and current density J they offer one of the simplest methods for finding $\boldsymbol{E}$ and $\boldsymbol{B}$ in electrodynamics. Interestingly, the proof we have carried out to confirm that the retarded potentials solve Maxwell's equations, applies equally well to the advanced potentials

$$\Phi(\boldsymbol{r},t) = \frac{1}{4\pi\epsilon_0}\int_V \frac{\rho(\boldsymbol{r'},t_A)}{s}\,dv' \tag{6.19}$$
$$\boldsymbol{A}(\boldsymbol{r},t) = \frac{\mu_0}{4\pi}\int_V \frac{\boldsymbol{J}(\boldsymbol{r'},t_A)}{s}\,dv' \tag{6.20}$$

where $t_A = t + \frac{s}{c}$ is the advanced time, and the advanced potentials also solve Maxwell's equations. This is a great curiosity for it corresponds to a signal being sent from the future about

the charge and the current densities, and is mathematically equally useful for solving Maxwell's equations. Only causality, the idea that only events from the past influence the present, leads to one working with the retarded potentials over the advanced potentials, mathematically they appear equally useful solutions.

Example 6.3. *A current is turned on suddenly at $t = 0$ in an infinitely long straight and electrically neutral wire lying on the z-axis, i.e.*

$$\boldsymbol{I}(t) = \begin{cases} 0\,\boldsymbol{e_z} & t \leq 0 \\ I\,\boldsymbol{e_z} & t > 0. \end{cases}$$

Find $\boldsymbol{E}$ and $\boldsymbol{B}$.

Due to the symmetry of the problem we may work with cylindrical coordinates and aim to find the fields at a distance r from the wire in the plane $z = 0$, i.e. $\boldsymbol{r} = \begin{pmatrix} r\cos\theta \\ r\sin\theta \\ 0 \end{pmatrix}$. Each element of current lies at $\boldsymbol{r'} = \begin{pmatrix} 0 \\ 0 \\ z' \end{pmatrix}$ for some z', hence $s = \sqrt{r^2 + z'^2}$. Now we may begin to compute the retarded potentials using equations (6.17) *and* (6.18).

Since the wire has no net charge then $\Phi = 0$. The retarded magnetic potential is

$$\boldsymbol{A}(r,t) = \frac{\mu_0}{4\pi}\int_{z'=-\infty}^{\infty} \frac{\boldsymbol{I}(t - \frac{s}{c})}{s} dz'.$$

Now when $t - \frac{s}{c} \leq 0$ then $\boldsymbol{I} = \boldsymbol{0}$, so we will not have contributions for the current for which

$$ct \leq s = \sqrt{r^2 + z'^2}$$

rearranging we find there is no contribution to the current integral for

$$c^2t^2 - r^2 \leq z'^2.$$

To rephrase, conversely we only have non-zero current in the integral when $z'^2 \leq c^2t^2 - r^2$ which implies that $|z'| \leq \sqrt{(ct)^2 - r^2}$. This is the limited part of the wire that the point a distance r away from the wire can communicate with via signals sent at the speed of light. Our integral

simplifies to

$$\begin{aligned}
\boldsymbol{A}(r,t) &= \frac{\mu_0 I}{4\pi}\int_{z'=-\sqrt{(ct)^2-r^2}}^{\sqrt{(ct)^2-r^2}} \frac{1}{s}dz'\,\boldsymbol{e_z} \\
&= 2\frac{\mu_0 I}{4\pi}\int_{z'=0}^{\sqrt{(ct)^2-r^2}} \frac{1}{\sqrt{r^2+z'^2}}dz'\,\boldsymbol{e_z} \\
&= \frac{\mu_0 I}{2\pi}\int_{z'=0}^{\sqrt{(ct)^2-r^2}} \frac{d}{dz'}\Big(\ln(\sqrt{r^2+z'^2}+z')\Big)dz'\,\boldsymbol{e_z} \\
&= \frac{\mu_0 I}{2\pi}\ln\Big(\frac{ct+\sqrt{(ct)^2-r^2}}{r}\Big)\,\boldsymbol{e_z}.
\end{aligned}$$

Now we can find the electric and magnetic fields (remembering that $\Phi = 0$):

$$\begin{aligned}
\boldsymbol{E}(r,t) &= -\boldsymbol{\nabla}\Phi - \frac{\partial \boldsymbol{A}}{\partial t} \\
&= -\frac{\mu_0 I}{2\pi}\frac{\partial}{\partial t}\Big[\ln\Big(\frac{ct+\sqrt{(ct)^2-r^2}}{r}\Big)\,\boldsymbol{e_z}\Big] \\
&= -\frac{\mu_0 I}{2\pi}\frac{\partial}{\partial t}\Big[\ln\Big(ct+\sqrt{(ct)^2-r^2}\Big)\,\boldsymbol{e_z}\Big] \\
&= -\frac{\mu_0 I}{2\pi}\Big(\frac{c}{\sqrt{(ct)^2-r^2}}\Big)\,\boldsymbol{e_z}
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{B}(r,t) &= \boldsymbol{\nabla}\times\boldsymbol{A} \\
&= -\frac{\mu_0 I}{2\pi}\frac{\partial}{\partial r}\Big[\ln\Big(\frac{ct+\sqrt{(ct)^2-r^2}}{r}\Big)\Big]\,\boldsymbol{e_\theta} \\
&= \frac{\mu_0 I}{2\pi r}\frac{ct}{\sqrt{(ct)^2-r^2}}\,\boldsymbol{e_\theta}
\end{aligned}$$

For $t \to \infty$ (the effects of the changing current are now infinitely far in the past) we recover the fields for the static case of the infinitely long straight wire of constant current $\boldsymbol{I} = I\,\boldsymbol{e_z}$.

7. Special Relativity

> 'What led me more or less directly to the special theory of relativity was the conviction that the electromotive force acting on a body in motion in a magnetic field was nothing else but an electric field.'
>
> Albert Einstein, in a letter to the Michelson Commemorative Meeting of the Cleveland Physics Society, 1952

Maxwell's equations inform us that electromagnetic waves travel at the speed of light $c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}$ which is a constant. This idea that a speed is constant regardless of the frame of reference is in conflict with the assumptions of relativity inherent in Newtonian mechanics, namely that the laws of physics hold equally well in any frame of reference which is moving at a constant velocity with respect to any other frame of observation. This contradiction between the two dominant areas of physics in the 19^{th} century was not immediately appreciated.

The long experience and success of Newtonian mechanics in which dynamics is explained via physical exchanges of energy, led physicists to presume that the electromagnetic wave must be propagating through some medium, called the luminiferous aether, against which the speed of light was determined. Famous experiments to detect the relative speed of electromagnetic waves when it moved in different directions through water such as the Fizeau experiment and then through the aether in the Michelson and Morley experiment failed to detect any change in the speed of light as it moved in different directions[1]. The possibility that the motion of the Earth dragged the aether around with it, so that no relative motion of the aether could be detected was then a possibility. Indeed to explain the results of the Michelson and Morley experiment

[1]Both experiments used interferometers to try to measure a difference in the speed of light as it was split up and travelled two different paths and then recombined, should the speed through the paths differ then interference patterns were expected to be detected. Fizeau's experiment in 1851 predated the notion of the electromagnetic wave and was concerned with measuring the speed of light through flowing water, while Michelson and Morley's experiments in 1887 tested whether the motion of the aether effected the electromagnetic waves. Both gave negative results.

for a fixed aether hypothesis, Fitzgerald and also Lorentz hypothesised that aether pressure caused the contraction of objects - since objects are held together using the electromagnetic force any impact on electromagnetic waves would impact on the length of physical objects - this was called the Lorentz-Fitzgerald contraction. However the dragged aether model was inconsistent with the long-known phenomenon of stellar aberration. The reluctance to give up the notion of a mechanical aether, with hindsight, can be said to be rooted in the Newtonian understanding of energy and mass. It would take a revolutionary reinterpretation of mechanics to obtain a picture that was consistent with both Newtonian mechanics (at low velocities) and with electromagnetism.

In 1905 Einstein published four papers which each changed the world. In the first he established that energy occurs in discrete quanta, which since the work of Max Planck had been thought to be a property of the energy transfer mechanism rather than energy itself - this work really opened the door for the development of quantum mechanics. In his second paper Einstein used an analysis of brownian motion to establish the physical existence of atoms. In his third and fourth papers he set out the special theory of relativity[2] and derived the most famous equation in physics, if not mathematics, relating energy to rest mass $E = mc^2$. Hence 1905 is often referred to as Einstein's annus mirabilis.

In this chapter we will give a brief description of the special theory of relativity. One consequence of relativity is that time and space are put on equal footing and we will need to develop the fundamental objects used in mechanics, such as time, mass, velocity and momentum so that they respect the four-dimensional space-time. While concepts from Newtonian mechanics will need to be developed, we will see that Maxwell's equations and so electromagnetism already respect the symmetries of space-time.

7.1 The Postulates of Special Relativity

The special theory of relativity was constructed from two postulates:

(1.) the laws of physics are independent of the inertial reference frame of the observer, and

(2.) the speed of light in a vacuum is a constant for all observers.

[2]The question of who first established the theory of relativity remains a matter of dispute, some facts which are not in doubt is that Lorentz had laid out the equations underlying the new mechanics before Einstein, and Poincaré had established his own principle of relativity before Einstein and had also derived $E = mc^2$, but as with many topics in mathematics and physics it is not always the first presentation of a new idea that takes the credit but often the clearest presentation of an idea. This is the case with Einstein's work. See `https://en.wikipedia.org/wiki/Relativity_priority_dispute`.

The first idea, which gives the theory its name, was not particularly revolutionary, even Newton only defined position and velocity up to measurement within an inertial frame, but the second idea that the speed of light was always constant, despite it being evident from Maxwell's equations, had ramifications for Newtonian mechanics. Newtonian mechanics was constructed on an absolute space, with time a parameter for the dynamics. Observations taken in different inertial frames (those whose coordinates only differed by a relative velocity) could be related by the Galilean transformations. Suppose that frame S' (in which the measured coordinates and speeds are denoted with a prime, e.g. t', x', y' and z') is moving with constant velocity[3] $\boldsymbol{v} = v\,\boldsymbol{e_x}$ with respect to a frame S (in which the measured coordinates and speeds are denoted without a prime, e.g. t, x, y and z), then the Galilean transformations are

$$\begin{aligned} t' &= t \\ x' &= x - vt \\ y' &= y \\ z' &= z. \end{aligned}$$

The Galilean transformations lead to a simple relation for the measurement of speed in the x-direction between the two frames. Consider a speed u_x in the x-direction measured in frame S, then the Galilean transformations tell us that the speed measured in frame S' is

$$u'_x = \frac{dx'}{dt'} = \frac{dx'}{dt} = u_x - v$$

where we see the straightforward addition of velocity required to transform measurements of velocity in the two frames. If a speed u'_x is reported in frame S', then the speed in the S frame is $u'_x + v$ because frame S' is moving with speed v relative to S in the $\boldsymbol{e_x}$ direction. So if an electromagnetic wave is moving at the speed of light c in S' then it is moving at the speed $c+v$ in S. The equations of electromagnetism would be valid in frame S' but not in frame S. By the first of Einstein's postulates this idea is ruled out, and by the second postulate the importance of the speed of light is singled out[4]. The Newtonian rules to measure velocities will need to be modified, and the Galilean transformations are not consistent with the postulates of special relativity.

The inhomogeneous electromagnetic wave equations $\Box\Phi = -\frac{\rho}{\epsilon_0}$ and $\Box\boldsymbol{A} = -\mu_0\boldsymbol{J}$ are not preserved under a Galilean transformation. We would find that $\frac{\partial}{\partial x} = \frac{\partial}{\partial x'}$, $\frac{\partial}{\partial y} = \frac{\partial}{\partial y'}$ and $\frac{\partial}{\partial z} = \frac{\partial}{\partial z'}$

[3]One might imagine this is a special choice, but given two frames whose relative velocity is $\boldsymbol{v}$ one can always choose a coordinate basis in each frame such that $\boldsymbol{e_x}$ and $\boldsymbol{e_{x'}}$ are in the direction of the relative velocity $\boldsymbol{v}$.

[4]One could certainly argue that only the first postulate is necessary, since the validity of electromagnetism in all frames of reference then guarantees the second postulate.

but $\frac{\partial}{\partial t} = \frac{\partial}{\partial t'} + \frac{\partial x'}{\partial t}\frac{\partial}{\partial x'} = \frac{\partial}{\partial t'} - v\frac{\partial}{\partial x'}$, so that the d'Alembertian operator transforms as

$$\Box'\Phi = -\frac{1}{c^2}\frac{\partial^2\Phi}{\partial t'^2} + \boldsymbol{\nabla}'^2\Phi = -\frac{1}{c^2}\Big(\frac{\partial}{\partial t} + v\frac{\partial}{\partial x}\Big)^2\Phi + \boldsymbol{\nabla}^2\Phi \neq \Box\Phi.$$

If we had a solution to $\Box\Phi = -\frac{\rho}{\epsilon_0}$ in one Galilean reference frame, then we would not have the same solution in another reference frame. The Galilean transformations do not respect the equations of electromagnetism.

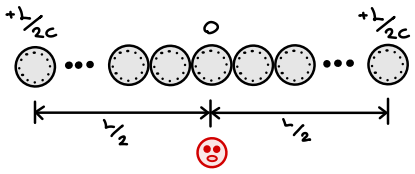
The Lorentz force law which depends upon the velocity of a test charge is also now seen to present us with some problems (if the Galilean transformations of an absolute space and time are maintained). Consider an infinite line of steady current directed in the $\boldsymbol{e_x}$ direction running in an infinite wire on the x-axis. It produces a circumferential static magnetic field $\boldsymbol{B} = \frac{\mu_0 I}{2\pi r}\boldsymbol{e_\theta}$ where r and θ are taken to stand for the radial coordinate and the angular coordinate in cylindrical coordinates organised about the x-axis. Now a test charge Q, a radial distance R from the wire moving with velocity $\boldsymbol{v} = v\,\boldsymbol{e_x}$ experiences a force $\boldsymbol{F} = Q(\boldsymbol{v}\times\boldsymbol{B}) = -Q\frac{\mu_0 I v}{2\pi R}\boldsymbol{e_z}$. However measured from the S' frame it is the wire that has a velocity $-\boldsymbol{v} = -v\boldsymbol{e_{x'}}$ and the particle is stationary - S' is called the rest-frame of Q. An observer in frame S' thinks that Q experiences no force due to the magnetic field of the current. This is very similar to the problem resolved by Faraday in deducing that a changing magnetic field produces an electric field, however in this example the magnetic field whose field lines form concentric cylinders about the wire are left unchanged by a Galilean transformation. This is another indication from electromagnetism that the Galilean transformations will not survive the postulates of relativity.

7.1.1 Simultaneity

The notion of an absolute time, such as appears in the Galilean transformations is problematic, and is in conflict with the constant speed of light. However observers in relative motion to each other may use the fact that the speed of light is the same no matter in which frame the light speed is being measured to deduce statements about the time as measured in the other frame. The constant nature of the speed of light becomes an absolute around which the measurement of time and distance can be based. Einstein reasoned that one could use light to synchronise clocks which are located in different places.

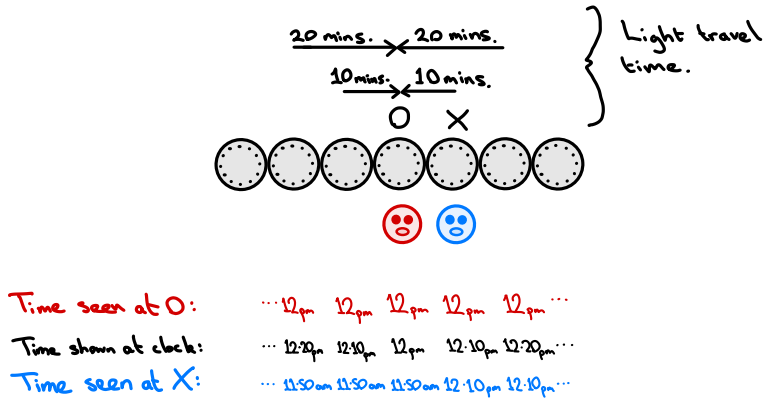
Let us think a little bit about synchronised clocks in a single frame of reference. Imagine a sequence of clocks densely laid out as far as they eye can see along the x-axis. Let us suppose that the clocks have been set so that viewed from the origin they all read the same time, and all tick in unison. For this to be the case the time on the clocks must be set up to account for the delay in a light signal sent from each clock to the origin. A clock at a distance $\frac{L}{2}$ from the

origin is read using light waves which take a time $\frac{L}{2c}$ to reach the observer at the origin. If the clock manages to show the same time as the clock at the origin then since the signal is showing the time in the past, it must be that the clock is running $\frac{L}{2c}$ ahead of the clock at origin.



Clocks runnning fast away from O appear synchronised at O

Clocks which appear synchronised at one point in a reference frame but which are not located at that point, we deduce need to run ahead. Locally the clocks all tell different times. If we consider the same set of clocks observed from another position in the same frame then the clocks will not appear synchronised.



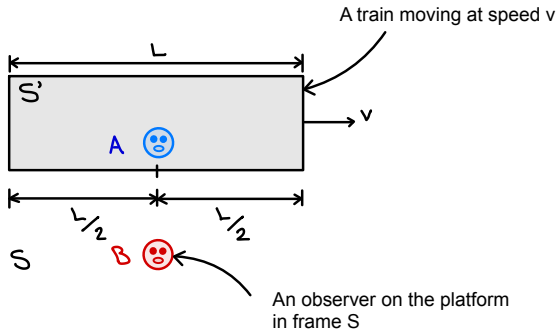
Clocks synchronised to read 12pm at O are not synchronised at X

One cannot then set up a system of clocks throughout space which would appear synchronised at all points in space. As it happens this notion of reading the same time, of actually seeing the same time on all clocks from one point will not be our aim in synchronising clocks. We have no problem computing the time it takes light to travel a distance so we can accommodate that delay. What we are interested in is being able to refer to *the* time in a frame of reference. One simple way to do this is to set two clocks to the same time at O and then slowly move a second clock to X, another way is to use light signals to synchronise clocks. With either of these methods, the light travel time means that at O the observer will see the clock reading $\frac{L}{c}$ in the past (if O and X are separated by a distance L). However within a frame we can take account of the time delay in reading the time from a distant clock due to the speed of light. When we talk of synchronised clocks in a frame this will be what is meant: that a collection

of synchronised clocks starting at O have been slowly moved throughout the frame. As viewed from O a clock a distance L away reads a time $\frac{L}{c}$ in the past. The usefulness of this arrangement is that if one records an event occurring at $\boldsymbol{r}$ in the frame, then one need only read off directly the time on the local clock at $\boldsymbol{r}$ to say when the event happened. This will be agreed by all observers (once they take account of the travel time for a light signal from $\boldsymbol{r}$) at different points in the frame, even though it will not be the time on their local clock. Notice that if the clocks are all synchronised in a frame then when that frame is put in motion, and all the clocks with it, the set of clocks will still be synchronised.

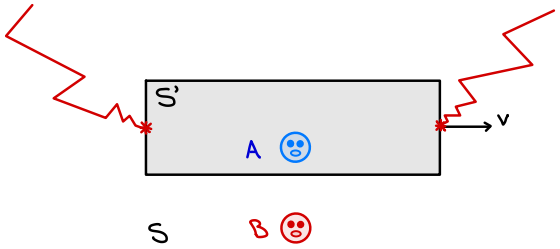
Still we must wonder how to go about practically synchronising distant clocks. Suppose we have clocks at two locations O and X, then the clock at at O measures the time at O while the clock at X measures the time at X. In order to determine a common time at both O and X one can use the constant speed of light: the time it takes a light signal to travel from O to X is the same as the time it takes a light signal to travel from X to O. One may send a light signal from O at $t = 0$ to X where it is reflected by a perfect mirror back to O and received at time $t = t_0$. Then the time at which the light reaches X is $\frac{t_0}{2}$. Synchronising clocks with this method would mean that at O the observer can ascribe the time $\frac{t_0}{2}$ to the clock at X at some time in the past. Alternatively a signal could be sent at the speed of light from the observer at O which when it reaches a distant clock is understood to mean that the distant clocks are set to an agreed time. For a light signal sent from the midpoint between two clocks, the two clocks will be set to read the same time when they receive the signal.

Now Einstein thought about the consequences for synchronised clocks in one frame as observed in another frame which is in motion with respect to the first. In a thought-experiment[5], or gedanken experiment, he wondered about two frames: one given by S' is the frame of a train-carriage, and a second S is the frame of a station through which the train passes at a constant speed v.

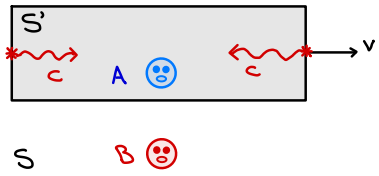


[5]Much of the material in this chapter follows the wonderful book 'Special Relativity for the Enthusiastic Beginner' by David Morin, you are encouraged to seek this book out for many more takes on wonderful puzzles and paradoxes, both old and new.

He imagined two lightning strikes occurring simultaneously in frame S (as measured by an observer B in the middle of the platform), striking the front and the back of the carriage, just as the middle of the train carriage, where observer A sits, is lined up with the observer B on the platform (so that the flash takes the same time to travel to the observer B from both the front and the back of the train).



Simultaneous lightning strikes as seen by B



The train moves towards the front flash and away from the rear flash as the light travels towards A

In this thought-experiment the train's motion is unaffected by the lightning strikes[6]. Einstein wonders what the passenger A, seated in the middle of the carriage observes. Since the speed of light is finite it takes a small, finite time for a flash to reach the passenger. In the time t the train has continued its motion at speed v, so that the passenger has moved closer by vt to the lightning strike at the front of the carriage and further from the lightning strike at the back. Suppose that it takes t_F for the front flash to reach passenger A, then the flash has to travel a distance

$$\frac{L}{2} - vt_F = ct_F$$

hence

$$t_F = \frac{L}{2(c+v)}.$$

On the other hand if it takes a time t_R for the flash to travel from the rear of the carriage to passenger A then in that time, as passenger A has moved a distance vt_R, then the flash from

[6]Thought trains[7]are more robust than real trains.

[7]Not to be confused with trains of thought which are fragile things.

the rear travels a distance to passenger A:

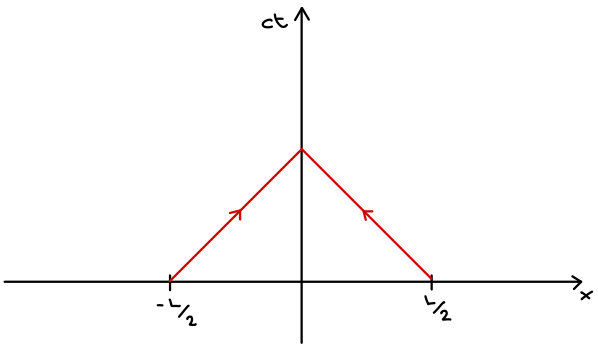
$$\frac{L}{2} + vt_R = ct_R$$

hence

$$t_R = \frac{L}{2(c-v)}.$$

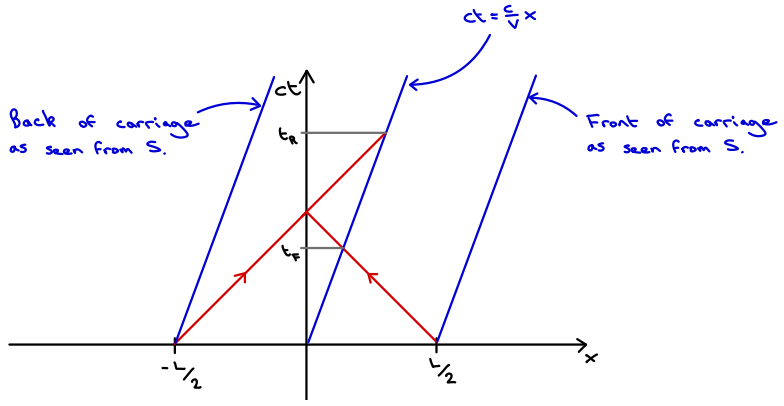
In other words, the distance that the flash from the front of the carriage has to travel is less than the distance the flash from the back of the carriage has to travel, because of the motion of the passenger A. Therefore the passenger will record that the lightning strike at the front of the carriage occurs before the lightning strike at the back of the carriage, essentially because the speed of light is constant. Since the two observers disagree about whether the two lightning strikes were simultaneous, the idea of an absolute time must be abandoned.

It is helpful to think about the same scenario in terms of a plot of ct against x as viewed from S:



Observer B is located at $x = 0$ and as time progress they stay on the line $x = 0$, since they do not move. The lightning strikes the carriage at its front at $x = \frac{L}{2}$ and at its back at $x = -\frac{L}{2}$. Light signals from these strikes travel in all directions at the peed of light, but we have just sketched those rays that go towards observer B at $x = 0$. Light rays are drawn at forty-five degrees to the horizontal, because they travel at speed c: in one second a light ray travels a distance c, and because our vertical axis is ct then $t = 1$ corresponds to c on the vertical axis. So light traverses one unit horizontally for every one unit vertically on this diagram. Of course it may travel in the positive or the negative x direction. The lines we have drawn are called worldlines, and they show what happens to an object as time progresses, for the light ray we may think about a wavefront for a plane wave (or one may prefer to think about a photon). The worldline of observer B is $x = 0$ and overlaps with the ct axis. On top of the diagram we

now sketch the worldlines for A, as well as the front and back of the train carriage as seen in S:



We may repeat the computation using the diagram, where we have shown against the axes of S the times t_F and t_R indicating when the flashes reach A from the front and rear respectively. The worldline of A is $ct = \frac{c}{v}x$, as seen by B. The slope of the worldline encodes the speed of A according to B. The worldline of the flash from the front is $ct = -x + \frac{L}{2}$. Let the event 'A sees the front flash' occur at (t_F, x_F). These coordinates satisfy the equations:

$$ct_F = \frac{c}{v}x_F$$
$$ct_F = -x_F + \frac{L}{2}$$

so we find that

$$t_F = \frac{L}{2(c+v)}$$

as before. The flash from the back of the carriage has a worldline given by $ct = x + \frac{L}{2}$ hence we must solve

$$ct_R = \frac{c}{v}x_R$$
$$ct_R = -x_R + \frac{L}{2}$$

for the event 'A sees the flash from the back' at (t_R, x_R). We obtain,

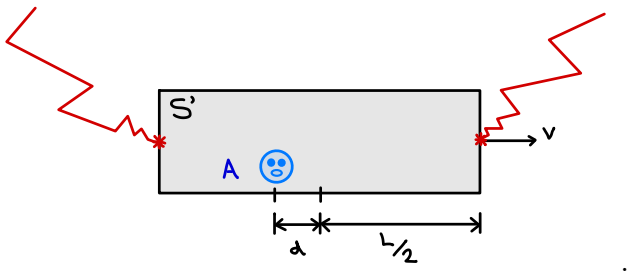
$$t_R = \frac{L}{2(c-v)}.$$

We obtain the same results as before and B deduces that A does not see the two lightning flashes simultaneously, even though B does.

In the thought-experiment we effectively used lightning strikes to synchronise clocks (that would be located on the front and back of the carriage) in the frame S, both lightning strikes

occur at the same time according to B, although they happen in different places and distant from B. However in frame S' the events are not occurring at the same time and so clocks in the moving frame are not synchronised using the lightning strikes. The use of light to communicate the signals is important to the problem, only in as much as the speed of light is a quantity that the two observers can agree on (due to the second postulate).

Let's think about another variation on this same thought-experiment. It is interesting to ask where on the train should passenger A sit if they were to see the two lightning flashes occurring at the same time? Our intuition tells us that they must sit behind the middle of the carriage towards the rear, since the train is moving towards the front flash and away from the rear flash this should even up the journey times for the two light flashes. So let us suppose that A sits a distance d behind the centre of the carriage:



Now we suppose that we picked d just right so that in the time t both the flash from the front of the carriage and the flash from the back of the carriage reach passenger A (in which time passenger A moves forwards vt according to B) then

$$(\frac{L}{2} + d) - vt = ct = (\frac{L}{2} - d) + vt.$$

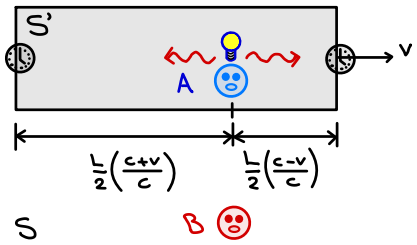
So we see that $d = vt$ (which we might have guessed) but also that $t = \frac{L}{2c}$, so that $d = \frac{vL}{2c}$. The position that A has to sit in splits the carriage in the ratio

$$\frac{\frac{L}{2} + d}{\frac{L}{2} - d} = \frac{\frac{L}{2} + \frac{vL}{2c}}{\frac{L}{2} - \frac{vL}{2c}} = \frac{c + v}{c - v}.$$

Notice that this is also the ratio $\frac{t_R}{t_F}$ for the different times it takes the light signals to reach A which we calculated above when A sits in the middle of the carriage.

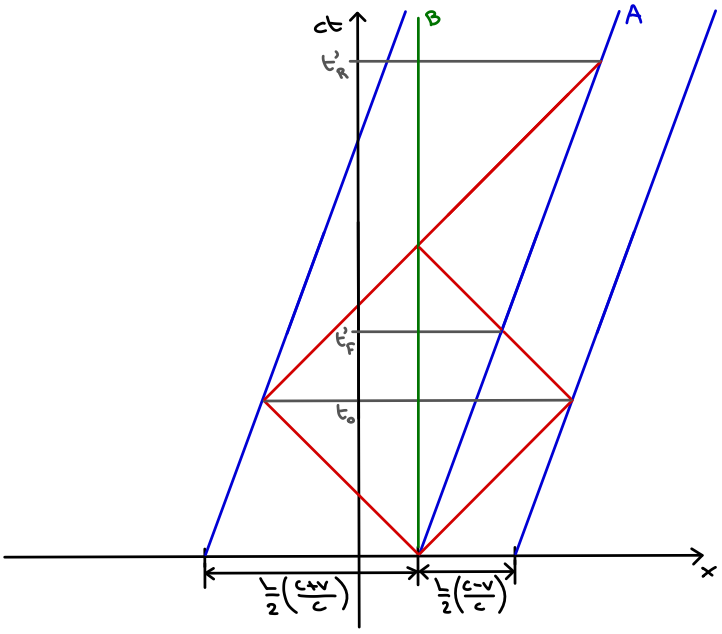
Finally we consider one more gedanken experiment in the previous setting of the train, but now we imagine there are two clocks which are synchronised with the S' frame time, the time on A's pocket-watch. We contrive a situation in which A sets off a flash gun sending out a light signal that reaches both clocks at the same time according to observer B, and B takes a photograph of the illuminated clock-faces. We realise that A must first sit in a new position

in the carriage before setting off their flash. For among the light waves travelling away from passenger A those that move forward are chasing the front of the train which is moving away at speed v, while those light waves that travel to the back of the train find that the rear of the train is moving at speed v towards them, all according to observer B. This is an inversion of the situation we considered earlier, and we realise that A must sit a distance $d = vt = \frac{vL}{2c}$ in front of the mid-point of the carriage. The location that A sits splits the carriage in the same ratio as before (just now the longer section is the rear section of the carriage).



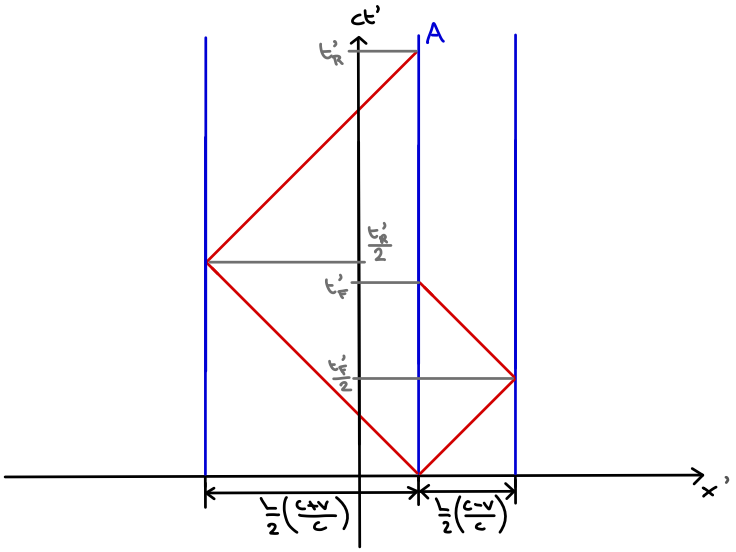
A sends a light signal that reaches the front and back of the carriage at the same time as measured by B in S

Now although for B the light hits the clocks simultaneously, for A the light will illuminate the front clock before the back clock (because it is closer to A). The clocks are synchronised for A and so A is able to deduce what time the clocks read when the flash illuminated them: this is what will be shown in B's photograph. When A sees each clock illuminated they record the times t'_F and t'_R for the front and rear clocks, and so A can deduce that the back clock is illuminated by the flash when it read $\frac{t'_R}{2}$, which is later than the time the front clock is illuminated, which A deduces to be $\frac{t'_F}{2}$. In the spacetime plot we sketch:



which shows that B sees the light emitted by A reach the front and the back of the carriage at the same time in their frame, but also shows the times that A records for the two different events. We also show the position of B and their camera throughout the experiment and the light rays for the flash as they return to A and B. What does B see in the photograph of the two clocks? A will tell us immediately that B will see a later time on the rear clock than on the front clock, and we can work out the time difference.

In the S' frame of A we sketch,



then if the flash was set off at time $t' = 0$ then A computes

$$\frac{t'_F}{2} = \frac{L}{2c^2}(c - v)$$
$$\frac{t'_R}{2} = \frac{L}{2c^2}(c + v)$$

and these are the times that B sees in the photographs. Even though they are synchronised in the frame S', in S the back clock reads a later time than the front clock. The time difference is

$$\Delta t' = \frac{t'_F}{2} - \frac{t'_R}{2} = -\frac{vL}{c^2}.$$

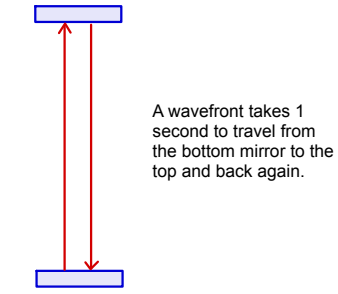
This is a measure of the loss of simultaneity between the frames and in any similar situation in which we will assume clocks in a moving frame have been synchronised, one may deduce that clocks towards the rear are running ahead of clocks towards the front.

7.1.2 Time Dilation and Length Contraction

Let us consider two further thought experiments to motivate modifications to the Galilean transformations. The first will demonstrate time dilation and the second the shortening of

length in a relatively moving frame.

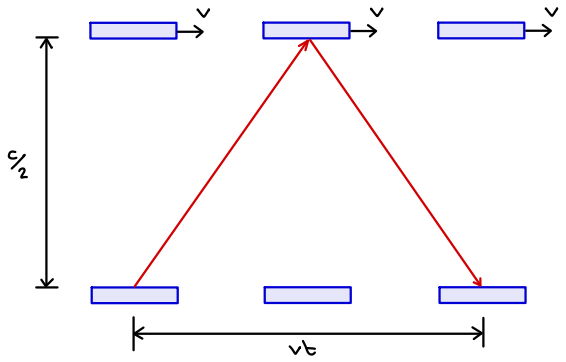
Consider a clock formed of two perfect mirrors separated vertically such that an electromagnetic wave bounces between the mirrors and takes one second to travel from the bottom mirror to the top mirror and back again. We can follow a specified wavefront of the electromagnetic wave in order to associate a location to the wave. It is consequently a very tall clock, it has height $h = \frac{c}{2}$ metres where c is the speed of light (hence $h \approx \frac{299792458}{2} = 149,896,229$ metres in a vacuum!).



A ticking light-clock

Let us construct two identical clocks of this type which 'tick' in unison and are perfectly synchronised. Let us set one light-clock in motion with a speed v in the x-direction and consider two observers: observer A moving with a light-clock in frame S' and observer B in frame S with their own clock, who sees A and their light-clock moving at speed v along the x-axis.

Suppose that the origins of the two frames coincide at $t = 0$ as measured in frame S, and that at this time both clocks are at the origin of frame S (and S' too, since the frames coincide at this moment). As observer A in frame S' moves off at speed v the observer B in frame S observes the "ticking" of the relatively moving light-clock slow down. Schematically we indicate a view of the moving clock as seen from frame S below:



Snapshots of a light-clock moving at speed v along the x-axis as seen from frame S through t seconds as measured in S

The wavefront in the moving clock now is seen to move along the hypotenuse of a right-angled

triangle as the clock moves horizontally at speed v. Since the wavefront travels at speed c then in S it is seen that the moving clock ticks at a slower rate, while in the frame S' the clock is still ticking at the same rate. Shown above is the path taken by the wavefront as seen by B in frame S for the time t that B perceives that A's clock takes to measure one second. The distance travelled by the wavefront (in frame S) is $2\sqrt{\frac{v^2t^2}{4}+\frac{c^2}{4}} = \sqrt{v^2t^2+c^2}$ and B measures this cycle to take t seconds. Since B knows the speed of the wavefront is c then they deduce that

$$ct = \sqrt{v^2t^2+c^2}$$

which may be rearranged to give

$$t = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} \equiv \gamma.$$

This is the time that t deduces it takes one second in frame S' to pass. Notice that we have defined the factor $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$ which for $v^2 < c^2$ satisfies $\gamma > 1$. One second in the moving frame takes more than one second as measured in the frame S. If B waited and watched t' seconds tick out on the moving frame S' then in frame S this takes $\gamma t'$ seconds, i.e.

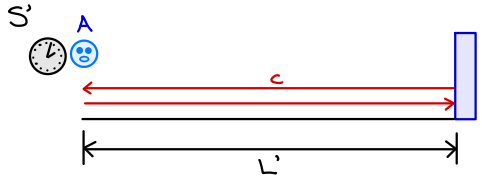
$$t = \frac{t'}{\sqrt{1-\frac{v^2}{c^2}}} = \gamma t'.$$

That time is seen to go more slowly in a relatively moving frame is known as time dilation. The effect is very small for human-scale speeds for a passenger aircraft (moving at roughly $600\,kmh^{-1}$) then $\gamma \approx 1.00000000001$, while if $v = \frac{c}{10}$ then $\gamma \approx 1.005$, and if $v = \frac{9c}{10}$ then $\gamma \approx 2.3$. In frame S' where A measures the clock ticking in the normal way, since A is not moving with respect to the light-clock, we may call the time change measured t' the proper time, and the time as measured from S we might call the observed time so that $t_{observed} = \gamma t_{proper}$. The proper time then is always the shortest measurement of the interval in any frame.

We may equally observe that as viewed by A, it is B who is moving at speed $-v$ along the x-axis. In which case A deduces that B's clock is running slow, in this case we would find that $t' = \gamma t$. This may seem contradictory but it is not. We just must bear in mind which frame is at rest relative to where the measurements are being made. In this case frame S' is at rest and A is measuring the time interval which we denote t' since A is in frame S'. It is therefore a potential source of confusion, where does the γ factor go? The key idea is that the stationary frame where the timing is being done does not pick up a γ with its measurement of the time interval. In practise one may simply pause and think which frame's time interval is dilated, and then put the γ in the appropriate place. A rule of thumb is that γ goes with the measurement in the frame in which the events being measured occur at the same location.

When S' is moving and B observes the wavefront then the events B measures are the start of the wavefront and then its return to the lower mirror and for B these two events occur in two different locations, while in frame S they occur in the same location hence $t = \gamma t'$, when S' is stationary then the events occur in frame S in the same location and then $t' = \gamma t$.

Let us consider a thought experiment that will show us that lengths contract in moving frames (as measured from a stationary frame). How can we measure length using the speed of light? Well we can use a setup like the light-clock turned on its side. We imagine that we construct a contraption, consisting of a straight rigid rod with a perfect mirror attached to one end (as drawn below), whose length is L' (the length is measured in S' the rest-frame of the rod).



The observer A will aim to measure its length using the wavefront of a light wave, whose arrival and departure time the observer measures accurately. The experiment will involve the wavefront traversing the length of the rod, being reflected by the perfect mirror and returning to its starting point. When conducted in the frame S' the light returns to its starting point in time $t' = \frac{2L'}{c}$. Hence A determines the length of the rod to be $\frac{ct'}{2}$.

Now we will change frames to S where B observes the experiment moving with speed v in the positive x direction. On the first leg of the journey the wavefront will take a longer time to reach the mirror, as the mirror is travelling away from the light. However on the return leg the observer A (at the end of the return leg) is moving towards the wavefront. We compute the time taken for each of the two legs t_1 and t_2, and note that viewed from S B supposes that the rod has length L:

$$ct_1 = L + vt_1 \implies t_1 = \frac{L}{c - v}$$
$$ct_2 = L - v_2 t_2 \implies t_2 = \frac{L}{c + v}$$

Hence B deduces that the experiment done by A to measure the rod takes a time

$$t_1 + t_2 = \left(\frac{L}{c - v} + \frac{L}{c + v} \right) = \frac{2Lc}{c^2 - v^2} = \frac{2L}{c}\gamma^2.$$

We may now use time dilation to relate the intervals measured in the two frames, so that B may compute the result of A's experiment for themselves. Now it is in frame S' that the events:

the start and the end of the wavefront's journey occur at the same location, at observer A. Consequently $(t_1 + t_2) = \gamma t'$ and so $\frac{c}{2}(t_1 + t_2) = \frac{c}{2}\gamma t' = \gamma L'$ and $\frac{c}{2}(t_1 + t_2) = \gamma^2 L$. Therefore,

$$L = \frac{1}{\gamma} L'.$$

As $\gamma \geq 1$ then $L \leq L'$. Length as measured from S has contracted in the moving frame S'. As for time dilation the effect is relative, not only is the length L' measured in S' seen as shortened in the direction of motion to $L = \frac{1}{\gamma}L'$ in S, but the result of the same experiment carried out in frame S to measure the length L is seen from frame S' to give the shortened length $L' = \frac{1}{\gamma}L$. As for time dilation this looks contradictory and one must take care to define which observations are being made in which frame. The frame in which the return of the light wave occurs at the same place (the frame of the experiment) is the frame in which the length of the rod is measured to be greatest, this length is called the proper length. A sense check when inserting the $\frac{1}{\gamma}$ factor is always recommended: it should go with the frame containing the length measuring experiment.

We have focussed on length contraction in the direction of motion of the frame, the longitudinal direction. Transverse to the relative motion there is no length contraction. Think of an experiment done in frame S' to collect mail for the train from the station. Suppose that beforehand A and B have measured the height of the train window (when the train is stationary) to be h_0. Next they will practise the mail collection with the moving train. A reaches out of the train window so as to collect the mail. In frame S B is not sure where A's hand will be and so separates out the mail and hangs letters at all heights along the z-axis, so as to guarantee the A collects one piece of mail. The question is which letter will be collected. If length in the z-direction is shortened by relative motion in the x-direction, then B sees that A collects the letter at height $h = \frac{1}{\gamma}h_0$ (by supposition). On the other hand, due to relativity, A sees B in motion and would see (if the z-direction exhibits length contraction) that the mail pole is shortened as $z' = \frac{1}{\gamma}z$ since A collects the mail at height $z' = h_0$ then it is the letter at height $z = \gamma h_0$ which is collected. With the assumption of transverse length contraction, A thinks that they collect a letter above the letter at h_0 (as measured on the platform), while B thinks that the letter collected lies below the letter at h_0. This experiment involves an event occurring at a single place and time, and relativity tells us that between inertial frames observers should agree on such measurements (although they may need to convert between the synchronised time and the origin of coordinates in each frame to do so). The only resolution to the contradiction is if there is no transverse length contraction. It is only in measuring the duration between two events or the distance between two events that the relative motion in a frame gives time dilation and length contraction. To elaborate, if lightning strikes the clock tower and stops the clock at 10.04pm, the time read on the clock-face is agreed by all observers, but observers

disagree about the time elapsed since the lightning strike (they compare two events then: the lighting strike, when they first look at their own watch and then some time later on looking again at their watch) - even so no one will dispute the time on the clock face even if they have spent all the time since the lightning strike travelling at 88mph.

These simple thought experiments changed the world and demonstrate the possibility that sometimes thought alone is able to outstrip intuition and experiment.

7.2 The Lorentz Transformations

In each of our thought experiments there were two events being contemplated (the two lightning strikes (or camera flashes), the electromagnetic wavefront starting and ending its journey in both the light-clock and the light-ruler), from which we could deduce the lack of simultaneity between frames, time dilation and length contraction. Let us suppose that the events have space separation $\Delta x'$ and time separation $\Delta t'$ as measured in S', and space separation Δx and time separation Δt as measured in S. We have emphasised three findings (where in each case the events contemplated are different):

- $\underline{\Delta t' = -\frac{v\Delta x'}{c^2}}$: From the lack of simultaneity, we have that $\Delta t = 0$ for the two camera flashes returning to B in S to give the photograph of the front and rear clocks on the train in our thought experiment, while in S', $\Delta t' = -\frac{Lv}{c^2} = -\frac{v}{c^2}\Delta x'$, where we note that since $\Delta x' = L$ then we must be measuring $x'_F - x'_R$ where x'_F and x'_R denote the coordinates of the front of the train carriage and the rear respectively as measured in S'. Consequently $\Delta t' = t'_F - t'_R = -\frac{v\Delta x'}{c^2}$ (we recall that the clock at the rear of the train runs ahead of the clock at the front of the train, hence the minus sign).

- $\underline{\Delta t = \gamma \Delta t'}$: From time dilation, where the two events are looking at a clock in he same location at two different times, so $\Delta x' = 0$ (the light-clock does not move relative to A).

- $\underline{\Delta x' = \frac{1}{\gamma}\Delta x}$: From length contraction, where the two events are the location of each end of a rod in frame S at the same moment in time in S' so $\Delta t' = 0$. This is not the same thought experiment that we used to deduce length contraction, but we can deduce from the way the length contraction is written what must have been measured. Remember the $\frac{1}{\gamma}$ factor goes with the proper length, so here we have written the length contraction for a length Δx (in S). The measurement of the two ends of a rod in S are carried out by an observer in S' hence $\Delta t' = 0$. When an observer deduces the length of an object they are able to say what the separation of the ends of the object is at the same moment in time. We write the expression in this way, because it will make the next steps simpler.

In addition we may add on a simple deduction about $\Delta x'$ between the two frames:

- $\underline{\Delta x' = -v\Delta t'}$ The relative change in position of the origin of frame S (so $\Delta x = 0$) as seen from S' over time.

These may not be the most natural way of framing each idea, but the presentation lends itself to deriving the Lorentz transformations.

The Lorentz transformations are expressions about the difference in measurements of events between two frames in relative motion. We have almost all the ingredients in the observations above, but in addition we can deduce that the Lorentz transformations, if they are to preserve physics between inertial reference frames, must map linear motion to linear motion between reference frames. The Lorentz transformations must be linear transformations, i.e.

$$\Delta t = A\Delta t' + B\Delta x'$$
$$\Delta x = C\Delta t' + D\Delta x'$$

where A, B, C and D are constants (remember v is constant). The coefficients A, B, C and D can't depend on particular coordinates otherwise the transformation between frames would depend on more than just their relative motion (which is what differentiates them), it also cannot be that the Lorentz transformations take special values at a particular point and time in spacetime - as the transformations must be relative and so cannot vary depending on the location of either frame.

From our earlier observations we can now deduce that[8]

$$\Delta t = \gamma\Delta t' + \gamma\frac{v}{c^2}\Delta x' = \gamma(\Delta t' + \frac{v}{c^2}\Delta x')$$
$$\Delta x = \gamma v\Delta t' + \gamma\Delta x' = \gamma(v\Delta t' + \Delta x').$$

Since $\boldsymbol{v} = v\,\boldsymbol{e_x}$ is in the x-direction there are no transformations on the measurements of intervals in y and z, so $\Delta y = \Delta y'$ and $\Delta z = \Delta z'$. Despite the transformations being related to two different measurements it is common to drop the Δ's: one can think of this as setting one of the events in S being measured to be at the origin (so $x = 0$) at time $t = 0$, and by arranging the frame S' so that its origins coincides with that of S at time $t' = 0$ then if the other event is at (t, x) and (t', x') as perceived from S in either frame, then $\Delta x = x$, $\Delta t = t$, $\Delta x' = x'$, and

[8]From time dilation, we obtain $A = \gamma$, then using simultaneity we find $B = \frac{v}{c^2}\gamma$, from length contraction we get $D = \gamma$, and from the relative motion of the frames we find $C = \gamma v$.

$\Delta t' = t'$. Altogether this gives a rewriting of the Lorentz transformations:

$$\begin{aligned} t &= \gamma(t' + \frac{v}{c^2}x') \\ x &= \gamma(vt' + x') \\ y &= y' \\ z &= z' \end{aligned}$$

It will be useful to work not with t but instead with ct, so that we have a fourth variable which has the dimension length, and also to use $\beta = \frac{v}{c}$, in this case the Lorentz transformations become

$$\begin{aligned} ct &= \gamma(ct' + \beta x') \\ x &= \gamma(\beta ct' + x') \\ y &= y' \\ z &= z' \end{aligned}$$

The above gives the relations between coordinates deduced in frame S for both S and S', if instead we were to make deductions from S', then while the primed coordinated would become unprimed, and vice versa, the relative speed of the frame being observed would also change from v to $-v$, so $\beta \to -\beta$ and the corresponding relations for measurement from S' are

$$\begin{aligned} ct' &= \gamma(ct - \beta x) \\ x' &= \gamma(-\beta ct + x) \\ y' &= y \\ z' &= z. \end{aligned}$$

The Lorentz transformations generalise the Galilean transformations and for $v \ll c$ then using $\beta = \frac{v}{c}$, $\frac{\beta}{c} \approx 0$ and $\gamma \approx 1$, we recover the Galilean transformations. Furthermore the Lorentz transformations preserve the speed of light in each frame, and leave Maxwell's equations invariant.

7.2.1 Addition of Velocity

The Lorentz transformations solve a problem that seemed to come from requiring that the speed of light is constant: how is it that one can fail to speed up light? If I am driving[9] at half of the speed of light and I turn on the car headlights, how fast do the light waves head off? We

[9] I do not drive and shouldn't be allowed to, not even at half the speed of light.

see that there is no problem in the rest frame of the car: the light travels at the speed of light, with respect to the car. The conundrum comes in a frame which see the car moving in the same direction as the light. The problem is resolved once we understand how to add velocities in relativity.

Consider the train (frame S') and station (frame S) of observers A and B again, and suppose that A bowls a cricket ball at speed u in the direction of the train's motion on the train. What speed does B on the platform deduce the ball is travelling at, given that the train is moving at speed v? Let us denote by frame S'' the rest frame of the cricket ball. We have a series of Lorentz transformations for moving between the frames and we will denote the gamma factor for the relative speed v by $\gamma(v)$ and for relative speed u by$\gamma(u)$,

$$\begin{aligned}
\Delta t &= \gamma(v)(\Delta t' + \frac{v}{c^2}\Delta x') \\
\Delta x &= \gamma(v)(v\Delta t' + \Delta x') \\
\Delta t' &= \gamma(u)(\Delta t'' + \frac{u}{c^2}\Delta x'') \\
\Delta x' &= \gamma(u)(u\Delta t'' + \Delta x'').
\end{aligned}$$

We may proceed in a number of ways, one method is to substitute $\Delta t'$ and $\Delta x'$ into Δt and Δx, and then read off the new Lorentz transformation that relates Δt to $\Delta t''$ - within which one can find the relativistic addition of speeds $u \oplus_R v$. However a simpler method is to choose $\Delta x'$ to be the distance the ball travels in time $\Delta t'$, then A reports that $\frac{\Delta x'}{\Delta t'} = u$. On the other hand, B reports the following

$$\begin{aligned}
\frac{\Delta x}{\Delta t} &= \frac{\gamma(v)(v\Delta t' + \Delta x')}{\gamma(v)(\Delta t' + \frac{v}{c^2}\Delta x')} \\
&= \frac{v + \frac{\Delta x'}{\Delta t'}}{1 + \frac{v}{c^2}\frac{\Delta x'}{\Delta t'}} \\
&= \frac{v+u}{1+\frac{uv}{c^2}}.
\end{aligned}$$

The relativistic addition of speeds is

$$u \oplus_R v = \frac{v+u}{1+\frac{uv}{c^2}}.$$

If one were travelling in a car at speed $v = \frac{c}{2}$ and turned on the headlights $u = c$ then a pedestrian would observe the wavefront of the electromagnetic wave travelling at speed $\frac{\frac{c}{2}+c}{1+\frac{c^2}{2c^2}} = c$ as expected. Furthermore $\lim_{u\to c}[\frac{v+u}{1+\frac{uv}{c^2}}] = c$. The velocity addition rule demonstrates that one cannot add speeds less than c to get to the speed of light, nor can one add speeds $v \leq c$ to obtain a speed greater than the speed of light.

7.3 Minkowski Space

Thinking about the Lorentz transformation in the previous section may put one in mind of a rotation in the (ct, x) plane, as there are some similarities at first glance, but a rotation would preserve the length $c^2t^2+x^2$, but here we find that $c^2t^2+x^2 = \gamma^2((1+\beta^2)t'^2+4vt'x'+(1+\beta^2)x'^2)$. If we had found coefficients of $(1-\beta^2)$ these would have cancelled with γ^2 and if we could have removed the cross-term in $t'x'$ this would have been fruitful. We may fix both problems by instead considering:

$$\begin{aligned}
-(ct)^2 + x^2 &= -\gamma^2(ct' + \beta x')^2 + \gamma^2(\beta ct' + x')^2 \\
&= -\gamma^2((ct')^2 + \beta^2 x'^2 + 2c\beta t'x') + \gamma^2(\beta^2(ct')^2 + x'^2 + 2c\beta t'x') \\
&= -\gamma^2((1-\beta^2)(ct')^2 - (1-\beta^2)x'^2) \\
&= -(ct')^2 + x'^2
\end{aligned}$$

we have learnt that $-(ct)^2 + x^2$ is a quantity left invariant by the Lorentz transformation above. Suppose one could evaluate this quantity and found it to equal to some positive value, for example $-(ct)^2 + x^2 = K^2$ are points on a hyperbola in the spacetime diagram, and the Lorentz transformation above is a translation along this hyperbola - let's see why. If we change variables so that $\gamma = \cosh\alpha$, note that this is well-defined since $\gamma \geq 1$ as is $\cosh\alpha$. Noting that $\frac{1}{\gamma^2} = 1-\beta^2$ we see that $\beta = \tanh(\alpha)$ and we have

$$\begin{pmatrix} ct \\ x \end{pmatrix} = \begin{pmatrix} (\cosh\alpha)ct' + (\sinh\alpha)x' \\ (\sinh\alpha)ct' + (\cosh\alpha)x' \end{pmatrix} = \begin{pmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} ct' \\ x' \end{pmatrix}$$

To see this corresponds to a translation along the hyperbola $-(ct')^2 + x'^2 = K^2$, we may write $ct' = K\sinh\alpha_0$ and $x' = K\cosh\alpha_0$ then

$$\begin{pmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} K\sinh\alpha_0 \\ K\cosh\alpha_0 \end{pmatrix} = \begin{pmatrix} K\sinh(\alpha_0+\alpha) \\ K\cosh(\alpha_0+\alpha) \end{pmatrix}$$

where we have used the hyperbolic double-angle formulae: $\cosh\alpha\sinh\alpha_0 + \sinh\alpha\cosh\alpha_0 = \sinh(\alpha_0+\alpha)$ and $\sinh\alpha\sinh\alpha_0 + \cosh\alpha\cosh\alpha_0 = \cosh(\alpha_0+\alpha)$. A translation along a hyperbola is the hyperbolic equivalent of translating a point around a circle in $\mathbb{R}^2$, which is just a rotation about the origin. Lorentz boosts are hyperbolic rotations.

Since y and z are unaffected by the Lorentz transformation above we may deduce that the quantity

$$-(ct)^2 + x^2 + y^2 + z^2 = -c^2\tau^2$$

is left invariant as well. The quantity τ is called the proper time[10] : recall that this is the time measured on the watch of the observer, for whom $\Delta x = \Delta y = \Delta z = 0$, or in the notation above $x = y = z = 0$, which explains the nomenclature. Once we allow variations in space, the proper time is the time measured on the observers watch as they traverse their world-line.

We have so far only considered a Lorentz transformation induced by two frames in relative motion in the x-direction, this is called a Lorentz boost in the x-direction. We may also consider Lorentz boosts which mix the y coordinate or the z coordinate and ct (corresponding to a relative motion in the y-direction or the z-direction between the reference frames). Such a transformation also leaves $-c^2 d\tau^2$ unchanged. We may construct a group of transformations, the Lorentz group, which leave the quantity $-c^2 d\tau^2$ unchanged, in much the same way that the orthogonal group $O(3)$ leaves the Euclidean distance $x^2 + y^2 + z^2$ invariant.

To construct the group we will define the Minkowski metric[11,12] on a four-dimensional space time,

$$\eta \equiv \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

which is an object whose components are $\eta_{\mu\nu}$ where $\mu, \nu \in \{0, 1, 2, 3\}$, and the array above contains $\eta_{\mu\nu}$ in the entry in the μ'th row and ν'th column. So that $\eta_{00} = -1$, $\eta_{11} = \eta_{22} = \eta_{33} = 1$ with all other components of $\eta_{\mu\nu}$ being zero. From this we may define four-vectors in $\mathbb{R}^4$ given by

$$\boldsymbol{Y} = \begin{pmatrix} Y^0 \\ Y^1 \\ Y^2 \\ Y^3 \end{pmatrix} \quad \text{and} \quad \boldsymbol{X} = \begin{pmatrix} X^0 \\ X^1 \\ X^2 \\ X^3 \end{pmatrix}$$

whose notation in components are Y^μ and X^μ and equip the space with the Minkowski inner

[10] We have opted to work with a mostly plus signature, but one could work equally well with $(ct)^2 - x^2 - y^2 - z^2 = c^2\tau^2$ as an invariant, doing so is referred to as working with a mostly minus signature.

[11] We commence the abuse of our familiar mathematical definitions here as the Minkowski metric is not positive definite as is implied by the definition of a metric, similarly the Minkowski inner product is also not positive definite but the constructions of both Minkowski inner product and Minkowski metric are close enough to the standard definitions that the misnomers have remained, and the lack of vocabulary will not confuse our work. Properly Minkowski space is a pseudo-Riemannian manifold in contrast to Euclidean space equipped with the standard metric which is a Riemannian manifold.

[12] This is also discussed in appendix A, where the emphasis is on introducing the index notation and the Einstein summation convention.

product given by

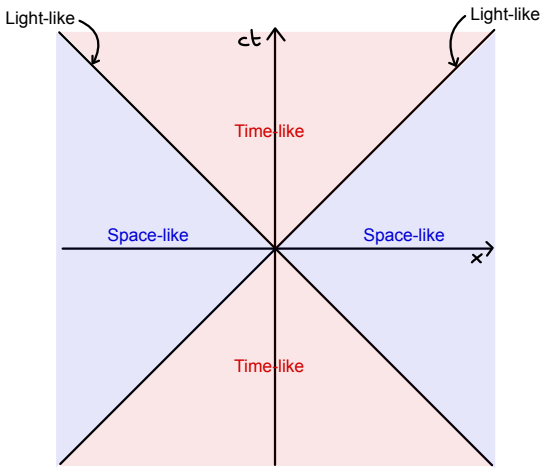
$$\begin{aligned}\boldsymbol{X}\cdot\boldsymbol{Y} &= \eta_{\mu\nu}X^\mu Y^\nu \\ &= -X^0Y^0 + X^1Y^1 + X^2Y^2 + X^3Y^3 \\ &= X_\mu Y^\mu\end{aligned}$$

where we have defined $X_\mu = \eta_{\mu\nu}X^\nu$ so that a contraction with the Minkowski metric tensor is similar to taking the transpose of a vector in $\mathbb{R}^3$, although it introduces a minus signs as $X_0 = -X^0$, $X_1 = X^1$, $X_2 = X^2$, and $X_3 = X^3$. The space $\mathbb{R}^4$ equipped with the Minkowski metric is often denoted $\mathbb{R}^{1,3}$ and called Minkowski space.

Four-vectors in Minkowski space no longer have a positive definite length-squared. All four-vectors fall into one of three classes:

$$S^2 = \boldsymbol{S}\cdot\boldsymbol{S}\begin{cases} < 0 & \boldsymbol{S} \text{ is called time-like} \\ = 0 & \boldsymbol{S} \text{ is called light-like or null} \\ > 0 & \boldsymbol{S} \text{ is called space-like.}\end{cases}.$$

Consider the subspace of $\mathbb{R}^{1,3}$ consisting of the $S^0 = ct$ and the $S^1 = x$ axes. Vectors in this two-dimensional sub-space are labelled by points which lie in one of, or at the meeting points of, the four sectors indicated below:



Let

$$\boldsymbol{S} = \begin{pmatrix} S^0 \\ S^1 \\ 0 \\ 0 \end{pmatrix}$$

be an arbitrary vector in $\mathbb{R}^{1,3}$ also lying entirely within $\mathbb{R}^{1,1}$ due to the zeroes in the third and fourth components, which denotes the spacetime location of a particle. So

$$\boldsymbol{S} \cdot \boldsymbol{S} = -(S^0)^2 + (S^1)^2$$

and hence if

$$\begin{aligned} (S^0)^2 > (S^1)^2 \qquad & \boldsymbol{S} \text{ is time-like.} \\ (S^0)^2 = (S^1)^2 \qquad & \boldsymbol{S} \text{ is light-like or null.} \\ (S^0)^2 < (S^1)^2 \qquad & \boldsymbol{S} \text{ is space-like.} \end{aligned}$$

If we plot the graph of motion of the x coordinate of $\boldsymbol{S}$ against ct the resulting curve is called the world-line of the event, and the notion of a word-line generalises similarly for more general displacement vectors. The gradient of the world-line

$$\text{Gradient} = \frac{\Delta(ct)}{\Delta(x)} = \frac{c}{v} \tag{7.1}$$

where v is the speed of the coordinate in the x direction. Hence if the coordinate moves at the speed of light, $\pm c$, then the gradient of the world-line is ± 1. In this case, when $x = vt = \pm ct$ (recalling that the particle is only moving in the x direction) then

$$\boldsymbol{S}^2 = -(ct)^2 + x^2 = 0 \tag{7.2}$$

so $\boldsymbol{S}$ is a light-like or null vector. If the gradient of the world-line is greater than one then $v < c$ and $\boldsymbol{S}$ is time-like, otherwise if the gradient is less than one then $v > c$ and $\boldsymbol{S}$ is a space-like vector. One of the consequences of the special theory of relativity is that objects cannot cross the light speed barrier and objects with non-zero rest-mass cannot be accelerated to the speed of light.

The spacetime at each point is split into four pieces. In the sketch above the set of null vectors form the boundaries of the light-cone for the origin. Given any arbitrary point in spacetime $\boldsymbol{P}$ the set of vectors $\boldsymbol{S} - \boldsymbol{P}$ are all either time-like, space-like or null. In the diagram above this would correspond to shifting the origin to the point $\boldsymbol{P}$, with spacetime again split into four pieces and their boundaries. The points which are connected to $\boldsymbol{P}$ by a time-like vector lie in the future or past light-cone of $\boldsymbol{P}$, those connected by a null vector lie on the surface light-cone of $\boldsymbol{P}$ and those connected by a space-like vector to $\boldsymbol{P}$ are outside the light-cone. With the supposition that nothing may cross the light speed barrier any point in spacetime can only exchange information with other points in spacetime which lie within or on its past or future light-cone.

In the two-dimensional spacetime that we have sketched it would be proper to refer to the forward or past light-triangle. The extension to four-dimensional spacetime is not easy to visualise. First consider extending the picture to a three-dimensional spacetime: add a second spatial axis y, as no spatial direction is singled out (there is a symmetry in the two spatial coordinates) the light-triangle of two-dimensions extends by rotating the light-triangle around the temporal axis into the y direction[13]. Rotating the light-triangle through three-dimensions gives the light-cone. The full picture for four-dimensional spacetime (being four-dimensional) is not possible to visualise and we refer still to the light-cone. However it is useful to be cautious when considering a drawing of a light cone and understand which dimensions (and how many) it really represents, e.g. a light-cone in four dimensions could be indicated by drawing a cone in three-dimensions with the implicit understanding that each point in the cone represents a two-dimensional space the drawing of which has been suppressed.

In all dimensions the light-cone is the cone at a point $\boldsymbol{P}$ traced out by all the light-like vectors connected to $\boldsymbol{P}$. No space-like separated points can exchange a signal since the message would have to travel at a speed exceeding that of light.

7.4 The Lorentz Group

Notice that the four-vector $\boldsymbol{S} = \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}$, which is the displacement four-vector in $\mathbb{R}^{1,3}$ (compared with the position vector $\boldsymbol{r}$ in $\mathbb{R}^3$ it has a time component) satisfies

$$S^2 = \boldsymbol{S} \cdot \boldsymbol{S} = -(ct)^2 + x^2 + y^2 + z^2 = -c^2\tau^2$$

This is a quantity that is invariant under the Lorentz transformations between frames with relative motion in the x, y or z directions. The Lorentz group is defined as the group of transformations which leave the Minkowski inner product invariant and will include the Lorentz boosts as well as other transformations, such as reflections and rotations. Let Λ be a matrix in $\mathbb{R}^{1,3}$ which transforms $\mathbb{R}^{1,3}$ such that the inner product between any two four-vectors is left

[13]By taking a slice of the three dimensional graph through ct and perpendicular to the xy-plane the two-dimensional light-triangle structure reappears.

invariant. Then if $\boldsymbol{X'} = \Lambda \boldsymbol{X}$ and $\boldsymbol{Y'} = \Lambda \boldsymbol{Y}$ then

$$\begin{aligned} \boldsymbol{X'} \cdot \boldsymbol{Y'} &= \eta_{\mu\nu} X'^{\mu} Y'^{\nu} \\ &= \eta_{\mu\nu} (\Lambda^{\mu}{}_{\rho} X^{\rho})(\Lambda^{\nu}{}_{\sigma} Y^{\sigma}) \\ &= \eta_{\lambda\kappa} (\Lambda^{\lambda}{}_{\mu} X^{\mu})(\Lambda^{\kappa}{}_{\nu} Y^{\nu}) \end{aligned}$$

where we have relabelled the dummy indices in the last line in order to make a direct comparison with

$$\boldsymbol{X} \cdot \boldsymbol{Y} = \eta_{\mu\nu} X^{\mu} Y^{\nu}.$$

If Λ leaves the Minkowski inner product invariant then it must satisfy

$$\eta_{\lambda\kappa} \Lambda^{\lambda}{}_{\mu} \Lambda^{\kappa}{}_{\nu} = \eta_{\mu\nu}$$

or in matrix notation,

$$\Lambda^T \eta \Lambda = \eta$$

where we treat η as a matrix. Notice that we need to use the transpose of Λ above because in $\eta_{\lambda\kappa}\Lambda^{\lambda}{}_{\mu}\Lambda^{\kappa}{}_{\nu} = \Lambda^{\lambda}{}_{\mu}\eta_{\lambda\kappa}\Lambda^{\kappa}{}_{\nu}$ when we treat η as a matrix then we see that the sum over κ in $\eta_{\lambda\kappa}\Lambda^{\kappa}{}_{\nu}$ encodes matrix multiplication, but the sum over λ in $\Lambda^{\lambda}{}_{\mu}\eta_{\lambda\kappa}$ does not (as λ runs over the column labels in η) , but since we may raise and lower pairs of dummy indices together then we can rewrite this as $\Lambda_{\lambda\mu}\eta^{\lambda}{}_{\kappa} = \Lambda^{T}{}_{\mu\lambda}\eta^{\lambda}{}_{\kappa}$ which is the matrix multiplication of Λ^T and η, where we treat η as a matrix.

Definition 7.4.1. *The Lorentz group $O(1,3)$ is defined by*

$$O(1,3) \equiv \{\Lambda \in GL(4,\mathbb{R}) \,|\, \Lambda^T \eta \Lambda = \eta,\ \eta = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\}$$

We may show that the transformations form a group:

(i) Closure: let $\Lambda_1, \Lambda_2 \in O(1,3)$ then

$$(\Lambda_1\Lambda_2)^T \eta \Lambda_1 \Lambda_2 = \Lambda_2^T (\Lambda_1^T \eta \Lambda_1) \Lambda_2 = \Lambda_2^T \eta \Lambda_2 = \eta$$

so $\Lambda_1\Lambda_2 \in O(1,3)$.

(ii) Associativity: Matrix multiplication is associative.

(iii) Identity: $\mathbb{I}$, the four-dimensional identity matrix, is in $O(1,3)$ as $\mathbb{I}^T \eta \mathbb{I} = \eta$.

(iv) Inverse: If $\Lambda \in O(1,3)$ then $\Lambda^T \eta \Lambda = \eta$ implies that $(\Lambda^T)^{-1}\eta\Lambda^{-1} = \eta$ and $(\Lambda^T)^{-1} = (\Lambda^{-1})^T$ is seen from taking the transpose of $\Lambda\Lambda^{-1} = \mathbb{I}$. Hence $\Lambda^{-1} \in O(1,3)$.

Among the transformations of the Lorentz group are the three Lorentz boosts (for frames in relative motion); three rotations (among the three spatial directions); three spatial reflections in the planes perpendicular to $\boldsymbol{e_x}$, $\boldsymbol{e_y}$ and $\boldsymbol{e_z}$; and a temporal inversion $t \to -t$. The subset of transformations which preserve the orientation of space, and the direction of time, are called the proper orthochronous, Lorentz transformations and form a subgroup denoted $SO^+(1,3)$, consisting of the rotations and Lorentz boosts (6 transformations) which are continuously connected to the identity element.

7.4.1 Lorentz Vectors and Tensors

A vector in $\mathbb{R}^3$ is not just a collection of three numbers but is an object that transforms, under a transformation of coordinates, in the same way as the vector $\boldsymbol{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$. A change of coordinates in $\mathbb{R}^3$ moves from one reference frame to another. In Minkowksi space we will therefore define a Lorentz vector as an object which transforms in the same way as the displacement four-vector $\boldsymbol{S} = \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}$ under a Lorentz transformation., i.e. as

$$\boldsymbol{S} \to \boldsymbol{S'} = \Lambda \boldsymbol{S}$$

which in components is

$$S^\mu \to S'^\mu = \Lambda^\mu{}_\nu S^\nu.$$

An object that transforms in this way is called a contravariant vector. We have also defined $S_\mu = \eta_{\mu\nu} S^\nu$, under a Lorentz transformation $\boldsymbol{S} \to \boldsymbol{S'} = \Lambda \boldsymbol{S}$ we have

$$S_\mu \to S'_\mu = \eta_{\mu\nu}\Lambda^\nu{}_\rho S^\rho = \Lambda_\mu{}^\rho S_\rho.$$

An object which transforms in this way is called a covariant vector. Rather than awkwardly talk about S_μ as being similar to, but different from the transpose of S^μ, we may talk directly about the contravariant vector component S^μ and the covariant vector component S_μ, or more simply covector (or, sometimes, it is called the dual vector). The two objects differ in the way they transform under a Lorentz transformation. Furthermore we may also introduce $\eta^{\mu\nu}$ which

raises indices as $S^\mu = \eta^{\mu\nu} S_\nu$. Then $S^2 = \eta_{\mu\nu} S^\mu S^\nu = \eta_{\mu\nu}\eta^{\mu\rho} S_\rho S^\nu = S_\nu S^\nu$. So $\eta_{\mu\nu}\eta^{\mu\rho} = \delta^\rho_\nu$. In terms of matrices $\eta^{\mu\nu}$ takes values which are the components of

$$\eta^{-1} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Covariant and contravariant vectors are dual to each other: we can take the inner product of a vector with a covector to find a scalar ($S^\mu T_\mu$ is a scalar). Consider a vector $\boldsymbol{V} = V^\mu \boldsymbol{e_\mu}$ where $\boldsymbol{e_\mu}$ form an orthonormal basis for the vector space, then a covector $\boldsymbol{U} = U_\mu \boldsymbol{e^\mu}$ where $\boldsymbol{e^\mu}$ form an orthonormal basis for the dual space, whose basis is defined by $\boldsymbol{e_\mu} \cdot \boldsymbol{e^\nu} = \delta^\nu_\mu$, as then $\boldsymbol{V}(\boldsymbol{e^\mu}) = \boldsymbol{V} \cdot \boldsymbol{e^\mu} = (V^\nu \boldsymbol{e_\nu}) \cdot \boldsymbol{e^\mu} = V^\mu$ which is a number, indeed it is a component of the vector $\boldsymbol{V}$. One may then define vectors by their action on the dual space V^*, and also define covectors (i.e. specify their components) by their action on the vector space V. We could define a vector abstractly as a map $\boldsymbol{V} : V^* \to \mathbb{R}$ by the inner product and similarly a covector by $\boldsymbol{U} : V \to \mathbb{R}$.

This idea of vector spaces and their duals allows us to generalise and construct objects whose components have more than one index, called tensors:

$$T : \underbrace{V^* \times V^* \ldots \times V^*}_{\text{m copies}} \times \underbrace{V \times V \ldots \times V}_{\text{n copies}} \to \mathbb{R}$$

which has a component with m contravariant indices and n covariant indices:

$$T^{\mu_1\mu_2\ldots\mu_m}{}_{\nu_1\nu_2\ldots\nu_n}$$

such an object is referred to as an (m, n) tensor. It may be defined by the way it transforms under a coordinate transformation, where each contravariant index transforms as a vector index, and each covariant index transforms as a covector index. I.e. under $\boldsymbol{S} \to \Lambda \boldsymbol{S}$ we have $T \to T'$ where in components

$$T'^{\mu_1\mu_2\ldots\mu_m}{}_{\nu_1\nu_2\ldots\nu_n} = \underbrace{\Lambda^{\mu_1}{}_{\rho_1}\Lambda^{\mu_2}{}_{\rho_2}\ldots\Lambda^{\mu_m}{}_{\rho_m}}_{\text{m copies}} \underbrace{\Lambda_{\nu_1}{}^{\sigma_1}\Lambda_{\nu_2}{}^{\sigma_2}\ldots\Lambda_{\nu_n}{}^{\sigma_n}}_{\text{n copies}} T^{\rho_1\rho_2\ldots\rho_m}{}_{\sigma_1\sigma_2\ldots\sigma_n}.$$

Tensors are a generalisation of the idea of a vectors and covectors (or dual vectors) and may be constructed as multilinear maps which act on products of V and V^*. Just like a vector they are defined by the way they transform under a coordinate transformation.

7.5 4-Displacement, 4-Velocity and 4-Momentum

The fundamental four-vector is the four-displacement[14] $\boldsymbol{S} = \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}$ which defines an event: it gives the location in time and space of an occurrence. By definition $\boldsymbol{S}$ is a Lorentz vector.

We can find other physically relevant four-vectors by combining the displacement four-vector $\boldsymbol{S}$ with Lorentz invariant quantities. For example the Lorentz four-velocity $\boldsymbol{V}$ (or proper velocity) is defined using the proper time, which is Lorentz invariant, rather than time t which is not. Starting from two events separated by Δt, Δx, Δy and Δz in frame S, but occurring at the same location in a frame S' moving at constant speed v between the two events. Now all observers agree on the proper time $\Delta\tau$ between the events. Then

$$\frac{\Delta \boldsymbol{S}}{\Delta\tau} = \frac{1}{\Delta\tau}\begin{pmatrix} c\Delta t \\ \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} = \frac{\Delta t}{\Delta\tau}\begin{pmatrix} c \\ \frac{\Delta x}{\Delta t} \\ \frac{\Delta y}{\Delta t} \\ \frac{\Delta z}{\Delta t} \end{pmatrix}$$

since the event is measured at the same place in the moving frame S', then it is the time measured in S' which is seen to be dilated: in that frame $\Delta t' = \Delta\tau$ so $\gamma\Delta\tau = \Delta t$. Now for an infinitesimal change $\Delta x = dx$, $\Delta y = dy$, and $\Delta z = dz$ observed in an infinitesimal time $\Delta t = dt$ then we may define the four velocity as

$$\boldsymbol{V} = \frac{d\boldsymbol{S}}{d\tau} = \gamma\begin{pmatrix} c \\ \frac{dx}{dt} \\ \frac{dy}{dt} \\ \frac{dz}{dt} \end{pmatrix} = \gamma\begin{pmatrix} c \\ v^1 \\ v^2 \\ v^3 \end{pmatrix}$$

where we see the appearance of the usual velocity for $v \ll c$ in the three spatial components of the four-vector $\boldsymbol{V}$, and we may write $\boldsymbol{V} = \begin{pmatrix} \gamma c \\ \gamma \boldsymbol{v} \end{pmatrix}$.

We can check that $\boldsymbol{V}^2$ is a Lorentz invariant quantity:

$$\boldsymbol{V}^2 = V^\mu V_\mu = \gamma^2(-c^2 + v^2) = -c^2\gamma^2(1 - \frac{v^2}{c^2}) = -c^2$$

[14] This is also known as the four-position.

The four-momentum is defined as $\boldsymbol{P} = m\boldsymbol{V}$ where m is the rest-mass (the mass as measured in the frame in which the mass is stationary).

$$\boldsymbol{P} = \begin{pmatrix} \gamma mc \\ \gamma m\boldsymbol{v} \end{pmatrix}$$

The spatial part of the four-momentum is the usual Newtonian momentum multiplied by γ, we will denote this quantity by $\boldsymbol{p} = \gamma m\boldsymbol{v}$ which is a relativistic momentum. The zeroth component is also intriguing:

$$P^0 = \gamma mc \approx mc + \frac{1}{2}\frac{1}{c}mv^2 + \dots$$

in the second term we recognise the usual kinetic energy multiplied by a factor $\frac{1}{c}$, this lead to Einstein's observation that $P^0 c$ is the relativistic energy (since terms in a sum will all be of the same type as the kinetic energy term, so the sum must be some form of energy):

$$E = P^0 c = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}} \approx mc^2 + \frac{1}{2}mv^2 + \dots .$$

This remarkable observation leads to the energy-momentum four vector being written in terms of E as $P = \begin{pmatrix} \frac{E}{c} \\ \boldsymbol{p} \end{pmatrix}$. The invariant quantity associated to $\boldsymbol{P}$ is

$$P^\mu P_\mu = -(\frac{E}{c})^2 + \boldsymbol{p}^2 = -m^2c^2$$

where we have used that $\boldsymbol{P}^2 = \gamma^2(-m^2c^2 + m^2v^2) = -m^2c^2$. Rearranging gives

$$E^2 = p^2c^2 + m^2c^4.$$

For a particle at rest we have $p = 0$ hence we find a particle's rest energy is

$$E = mc^2. \tag{7.3}$$

The equation also gives meaning to the relativistic energy for a massless particle ($m = 0$):

$$E = pc.$$

In electromagnetism we have been working with electromagnetic waves whose energy flux density is given by the Poynting vector (not to be confused in this chapter with the displacement four-vector) $\boldsymbol{S} = \frac{1}{\mu_0}(\boldsymbol{E} \times \boldsymbol{B})$. From the Poynting vector one can define the momentum density stored in the electromagnetic fields it is given by $\boldsymbol{g} = \frac{1}{c^2}\boldsymbol{S}$. For a monochromatic plane wave one can compute that the momentum $|\boldsymbol{g}| = \frac{U}{c}$ where U is the energy density of the plane wave. This

is expressing that $E = pc$ in the setting of electromagnetism, which was constructed without recourse to special relativity.

To solve dynamical problems in relativity, one uses the conservation of total energy where, for a particle of mass m moving with velocity $\boldsymbol{v}$, $E = \gamma(v)mc^2$, and conservation of total relativistic three-momentum, which, for a particle, is $\boldsymbol{p} = \gamma(v)m\boldsymbol{v}$ in any particular frame. That conservation of energy and momentum are valid for dynamics in some frame implies, by the first postulate of special relativity, that they hold in all Lorentz frames. However in transforming between frames the quantities may vary, after all a particle at rest in one frame has zero three-momentum, but in a frame in relative motion has a non-zero three-momentum. The law of physics (that energy and momentum are conserved in dynamics) is still valid in all frames, but the particular values of the quantities may vary. Since the four-momentum contains both the energy in its zeroth component and the momentum in the remaining components, the two conservation rules are equivalent to saying the total four-momentum is conserved in relativistic dynamics.

The four-momentum $\boldsymbol{P} = \begin{pmatrix} E/c \\ p_x \\ p_y \\ p_z \end{pmatrix}$ transforms as a Lorentz contravariant vector. Hence between two Lorentz frames S and S' with S' moving at speed v in $\boldsymbol{e_x}$ direction with respect to S we have

$$\begin{pmatrix} \frac{E}{c} \\ p_x \\ p_y \\ p_z \end{pmatrix} = \begin{pmatrix} \gamma(\frac{E'}{c} + \frac{v}{c}p'_x) \\ \gamma(\frac{v}{c^2}E' + p'_x) \\ p'_y \\ p'_z \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} \frac{E'}{c} \\ p'_x \\ p'_y \\ p'_z \end{pmatrix} = \begin{pmatrix} \gamma(\frac{E}{c} - \frac{v}{c}p_x) \\ \gamma(-\frac{v}{c^2}E + p_x) \\ p_y \\ p_z \end{pmatrix}.$$

The energy-momentum relation $E^2 = p^2c^2 + m^2c^4$ is invariant between Lorentz frames.

The relativistic version of Newton's second law is

$$\boldsymbol{F} = \frac{d\boldsymbol{p}}{dt}$$

where the above is a three-dimensional vector equation and $\boldsymbol{p}$ is the relativistic momentum, i.e. $\boldsymbol{p} = \gamma m\boldsymbol{v}$ where $\boldsymbol{v}$ is the velocity of the object whose rest mass is m. We should notice that this might seem like an unusual choice for defining the force, and that it might have been sensible to work instead with $\frac{d\boldsymbol{p}}{d\tau} = \boldsymbol{F_M}$, this is called the Minkowski force and differs the force defined above by a factor of $\frac{dt}{d\tau} = \gamma$, i.e. $\boldsymbol{F_M} = \gamma\boldsymbol{F}$. We may work with either force, so long as we are careful.

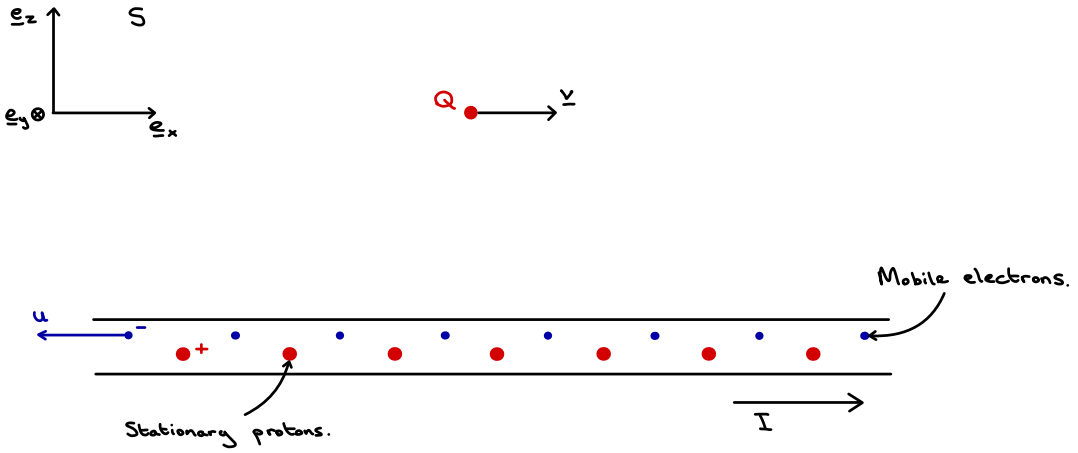
8. Relativistic Electrodynamics

Special Relativity is a necessary consequence of the need to reformulate mechanics so as to accommodate Maxwell's theory of electromagnetism. In this chapter we will investigate how electromagnetism is consistent with relativity, by seeing how the electric and magnetic fields transform under the Lorentz transformations and recasting Maxwell's equations in a manifestly Lorentz covariant form.

8.1 Lorentz Boosting the Magnetic Field of a Wire

In this section we will consider a simple example that was one of the ideas that motivated Einstein's work on special relativity. The idea is a simple one, the Lorentz force depends on the velocity $\boldsymbol{v}$ of the test charge, but what happens to this force in the rest-frame of test charge.

Consider an infinitely long wire, aligned along the x-axis, carrying a current $\boldsymbol{I} = I\,\boldsymbol{e_x}$. Now consider a test charge Q located in the plane $y = 0$ a distance z from the wire which is moving with a velocity $\boldsymbol{v} = v\,\boldsymbol{e_x}$:



The wire has zero net electric charge: the negative charges of the electrons are balanced by the

positive charges of the protons. The current is produced by electrons with velocity $\boldsymbol{u} = -u\,\boldsymbol{e_x}$, which gives a current $\boldsymbol{I} = \lambda u\,\boldsymbol{e_x}$ (we think of the wire as a line, hence the appearance of the line charge density λ of the electrons). The protons in the wire are stationary in the frame S. In light of relativity, we realise that the computation of the charge density as viewed from S must take account that distances in the rest-frame of the electrons will appear contracted in S. If the charge density of the electrons in their rest-frame is λ_0 then the charge density is increased as seen in S to $\lambda = \gamma(u)\lambda_0$. It increases because if the charge density in the rest-frame of the electrons is $\lambda_0 = \frac{Q}{L_0}$ then in S length contraction gives $L = \frac{1}{\gamma(u)}L_0$, while there is no effect on the charge due to the speed of the frame of reference, hence $\lambda = \frac{Q}{L} = \gamma(u)\frac{Q}{L_0} = \gamma(u)\lambda_0$. The charge density of the electrons is $\lambda_e = -\gamma(u)\lambda_0$ while the charge density of the protons is $\lambda_p = \gamma(u)\lambda_0$, to ensure that the total charge of the wire is zero.

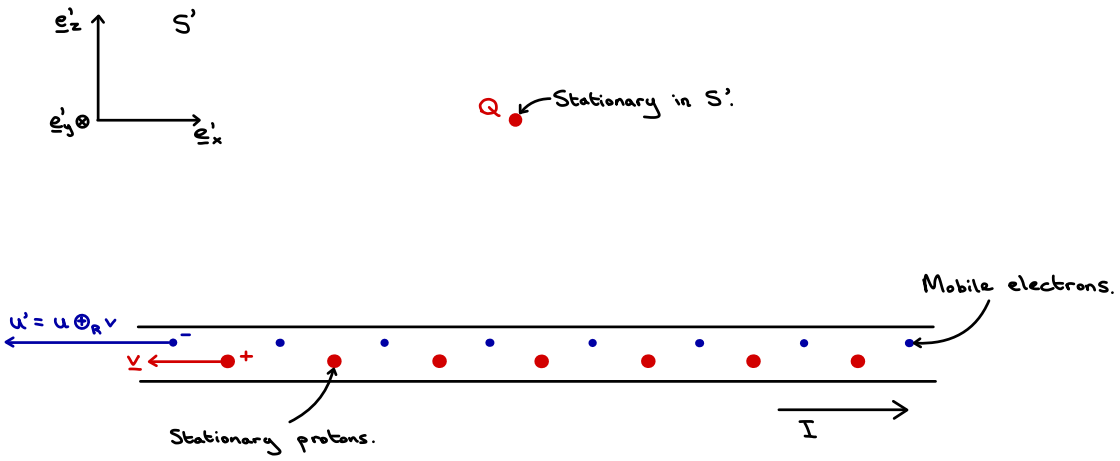
The infinite wire produces a magnetic field that circulates around the wire of size $|B| = \frac{\mu_0 I}{2\pi r}$ where r is the radial distance from the wire (as computed in example 5.2). At the location of the test charge (a distance z from the wire in the plane $y = 0$), the magnetic field is $\boldsymbol{B} = -\frac{\mu_0 I}{2\pi z}\,\boldsymbol{e_y}$. Then, in frame S, the Lorentz force law gives a force on the test charge

$$\begin{aligned}
\boldsymbol{F} &= Q(\boldsymbol{v} \times \boldsymbol{B}) \\
&= Q\Big(v\,\boldsymbol{e_x} \times -\frac{\mu_0 I}{2\pi z}\,\boldsymbol{e_y}\Big) \\
&= -Q\frac{\mu_0 v I}{2\pi z}\,\boldsymbol{e_z} \\
&= -Q\frac{\mu_0 \lambda_0 \gamma(u) u v}{2\pi z}\,\boldsymbol{e_z}
\end{aligned}$$

where we have substituted $I = \gamma(u)\lambda_0 u$. The test charge is subjected to a force that acts towards the wire. The force is due to the steady current in the wire and the motion of the test charge, so it is a magnetic force and a magnetic field. There is no electric field as there is no net source charge (the charge densities of the electrons and protons cancel each other out).

Now let us look at the same situation viewed from the rest-frame of the test charge, that is

from a frame S' moving at velocity $\boldsymbol{v} = v\,\boldsymbol{e_x}$ with respect to S:



The test charge Q is stationary in S', the electrons now have velocity $\boldsymbol{u'} = -u \oplus_R v\,\boldsymbol{e_x}$ where we indicate the relativistic addition of velocities by $\oplus_R$, and the protons move at velocity $\boldsymbol{v} = -v\,\boldsymbol{e_x}$. As the test charge Q is stationary then there is no magnetic force acting on Q in S', however we know that the laws of physics should be preserved between reference frames. It cannot be that the test charge is accelerated towards the wire in S but not in S'. What is the origin of the force on Q in S'?

The velocities of the charges in the wire has an impact on the charge densities due to length contraction. In S' we find

$$\lambda'_e = -\gamma(u')\lambda_0, \qquad \text{and} \qquad \lambda'_p = \gamma(v)\lambda_p = \gamma(v)\gamma(u)\lambda_0.$$

The lack of symmetry in the treatment of the proton and electron charge densities above is due to the fact that S is the rest-frame of the protons, but not of the electrons, but we may deduce the resting charge density of the protons in S because the wire has total charge density zero (i.e. we used $\lambda_e + \lambda_p = 0$ to write λ_p in terms of the charge density of the electrons in their rest-frame). We will now compute the total charge density in S': $\lambda'_e + \lambda'_p$, for which we will need $\gamma(u')$. We have,

$$u' = \frac{u+v}{1+\frac{uv}{c^2}}$$

so

$$\begin{aligned}\gamma(u')^2 &= \Big(1-\frac{u'^2}{c^2}\Big)^{-1}\\ &= \Big(1-\frac{1}{c^2}\Big(\frac{u+v}{1+\frac{uv}{c^2}}\Big)^2\Big)^{-1}\\ &= \Big(\frac{(c^2-u^2)(c^2-v^2)}{(c^2+uv)^2}\Big)^{-1}\\ &= \gamma(u)^2\gamma(v)^2\Big(1+\frac{uv}{c^2}\Big)^2. \end{aligned} \tag{8.1}$$

Consequently the total charge density in S' is

$$\begin{aligned}\lambda'_e+\lambda'_p &= \big(\gamma(u)\gamma(v)-\gamma(u')\big)\lambda_0\\ &= \gamma(u)\gamma(v)\Big(1-\big(1+\frac{uv}{c^2}\big)\Big)\lambda_0\\ &= -\gamma(u)\gamma(v)\frac{uv}{c^2}\lambda_0.\end{aligned}$$

The charge density on the wire is not zero and produces an electric field. The electric field due to an infinite wire of constant charge density was computed in example 3.2, and at the location of the test charge (a distance $z'=z$ from the wire):

$$\boldsymbol{E'}=\frac{\lambda'_e+\lambda'_p}{2\pi\epsilon_0 z}\,\boldsymbol{e_z}=-\gamma(u)\gamma(v)\frac{uv\lambda_0}{2\pi\epsilon_0 z c^2}\,\boldsymbol{e_z}$$

and using $c^2=\frac{1}{\epsilon_0\mu_0}$ we have the electric force on Q is

$$\boldsymbol{F'}=-\gamma(u)\gamma(v)\frac{Q\mu_0 uv\lambda_0}{2\pi z}\,\boldsymbol{e_z}.$$

This is almost, but not quite, the force experienced by Q in S due to the magnetic field. We must still relate the force in S to the force in S'. To do this we can use the Lorentz transformation between frames to relate the components of the force.

The force we use is defined by $\boldsymbol{F}=\frac{d\boldsymbol{p}}{dt}$, where $\boldsymbol{p}=\gamma(v)m\boldsymbol{v}$ for a particle of mass m moving at velocity $\boldsymbol{v}$. The Lorentz transformation relating S to S' is a boost by v in the $\boldsymbol{e_x}$ direction, and so mixes the temporal and the x-coordinates. The relativistic three-momentum transforms as part of the four-momentum, while time transforms as component of the four-displacement. Let us consider the components of the force:

$$F_x=\frac{dp_x}{dt}=\frac{\gamma(v)(\frac{v}{c}dP'^0+dp'_x)}{\gamma(v)(dt'+\frac{v}{c^2}dx')}=\frac{\frac{v}{c^2}\frac{dE'}{dt'}+\frac{dp'_x}{dt'}}{1+\frac{v}{c^2}\frac{dx'}{dt'}}$$

where we note that for Q in S' we have: $P'^0=\frac{E'}{c}$ is constant, so $\frac{dE'}{dt'}=0$ and Q is stationary in S', so $\frac{dx'}{dt'}=0$. Altogether we have (for our charge Q)

$$F_x=F'_x.$$

The y-component of the force transforms in a different way to the x-component:

$$F_y = \frac{dp_y}{dt} = \frac{dp'_y}{\gamma(v)(dt' + \frac{v}{c^2}dx')} = \frac{\frac{dp'_y}{dt'}}{\gamma(v)(1 + \frac{v}{c^2}\frac{dx'}{dt'})} = \frac{1}{\gamma(v)}F'_y$$

and the z-component transforms in the same way as the y-component:

$$F_z = \frac{dp_z}{dt} = \frac{dp'_z}{\gamma(v)(dt' + \frac{v}{c^2}dx')} = \frac{\frac{dp'_z}{dt'}}{\gamma(v)(1 + \frac{v}{c^2}\frac{dx'}{dt'})} = \frac{1}{\gamma(v)}F'_z.$$

In frame S' the only non-zero component of the force is the z-component, so we have

$$F_z = \frac{1}{\gamma(v)}F'_z = -Q\frac{\mu_0\lambda_0\gamma(u)uv}{2\pi z}$$

and $F'_x = F'_y = 0$. This is precisely the force due to the magnetic field in S, but in S' there is no magnetic field, instead the force is due to the electric field. This is remarkable, the balancing out of the forces comes from length contraction as well as the relativistic velocity addition and the change of coordinates between frames due to the Lorentz boost. All of this was implicitly encoded in Maxwell's equations.

8.2 The Lorentz Transformations of $\boldsymbol{E}$ and $\boldsymbol{B}$

We have seen that a magnetic field in one frame is an electric field in another frame. In general both the electric and the magnetic field will be non-zero and mixed by the Lorentz transformations.

Consider[1] two infinite planes of surface charge σ and $-\sigma$, forming the ideal capacitor in frame S. We studied the infinite plane of charge in example 3.3 and the ideal capacitor in problem 12 of the problem sets. Let the two planes of charge be parallel to $z = 0$, with the positive charge plane being $z = 0$ and the negative plane being $z = d$ where $d > 0$, then the electric field between the planes is

$$\boldsymbol{E} = \frac{\sigma}{\epsilon_0}\,\boldsymbol{e_z}. \tag{8.2}$$

Viewed from frame S', moving at velocity $\boldsymbol{v} = v\,\boldsymbol{e_x}$ as measured in S, the infinite planes of charge are moving with velocity $-\boldsymbol{v} = -v\,\boldsymbol{e_x}$. Consequently the charge density is increased due to the length contraction in the $\boldsymbol{e_x}$ direction, we have $\sigma' = \gamma(v)\sigma$, so the electric field becomes

$$\boldsymbol{E'} = \frac{\gamma(v)\sigma}{\epsilon_0}\,\boldsymbol{e_z} = \gamma(v)\boldsymbol{E}.$$

[1]We follow the approach of Griffiths in Introduction to Electrodynamics, and also Purcell's Electricity and Magnetism.

If we had boosted in the $\boldsymbol{e_y}$-direction by v we would have found the same transformation. However if we boost the ideal capacitor in the direction of the field, i.e. in the $\boldsymbol{e_z}$ direction, then the length contraction does not effect the charge density, only the distance d between the plates is length contracted which does not change the field. We may summarise this in the observation that a Lorentz transformation transforms the components of the field perpendicular to the boost. For a Lorentz boost in the $\boldsymbol{e_x}$ direction we have $E_x = E'_x$, $E_y = \gamma(v)E'_y$ and $E_z = \gamma(v)E'_z$, but this is not the whole story, we should adapt the setting to introduce a magnetic field as well as an electric field and see how they transform together.

In S' the moving planes of charge produce two surface currents $\boldsymbol{K}_{z=0} = -v\sigma'\,\boldsymbol{e_x}$ for the plane at $z = 0$ and $\boldsymbol{K}_{z=d} = v\sigma'\,\boldsymbol{e_x}$ for the plane at $z = d$. The magnetic field for an infinite plane of surface current was computed in example 5.4. The magnetic fields for the two planes cancel outside the capacitor, but within the planes they sum to give

$$\boldsymbol{B}' = \mu_0 K\,\boldsymbol{e_y} = \mu_0 v\sigma'\,\boldsymbol{e_y}.$$

This is a setting where we have non-zero components of both the electric and magnetic field:

$$E'_z = \frac{\sigma'}{\epsilon_0} = \gamma(v)\frac{\sigma}{\epsilon_0} \qquad \text{and} \qquad B'_y = \mu_0 v\sigma' = \mu_0 v\gamma(v)\sigma.$$

We may now consider these fields from another frame S'' moving with velocity $\boldsymbol{u} = u\,\boldsymbol{e_x}$ with respect to S'. The same considerations lead us to deduce that in S'' we have

$$E''_z = \frac{\sigma''}{\epsilon_0} = \gamma(w)\frac{\sigma}{\epsilon_0} \qquad \text{and} \qquad B''_y = \mu_0 w\sigma'' = \mu_0 w\gamma(w)\sigma.$$

where $w = u \oplus_R v = \frac{u+v}{1+\frac{uv}{c^2}}$. This is the most general result of Lorentz transforming both an electric and magnetic field in this setting, as subsequent Lorentz transformations in the $\boldsymbol{e_x}$ direction will lead to similar expressions.

We have computed $\gamma(w)$ in equation (8.1) (for $w = u'$), and so $\gamma(w) = \gamma(u)\gamma(v)\big(1 + \frac{uv}{c^2}\big)$. We find

$$\begin{aligned} E''_z &= \gamma(u)\gamma(v)\big(1 + \frac{uv}{c^2}\big)\frac{\sigma}{\epsilon_0} \\ &= \gamma(u)\big(E'_z + \frac{u}{c^2}\frac{B'_y}{\mu_0\epsilon_0}\big) \\ &= \gamma(u)\big(E'_z + uB'_y\big) \end{aligned}$$

and

$$\begin{aligned} B''_y &= \mu_0 w\gamma(u)\gamma(v)\big(1+\frac{uv}{c^2}\big)\sigma \\ &= \mu_0\Big(\frac{u+v}{1+\frac{uv}{c^2}}\Big)\gamma(u)\gamma(v)\big(1+\frac{uv}{c^2}\big)\sigma \\ &= \mu_0\gamma(u)\gamma(v)(u+v)\sigma \\ &= \gamma(u)(\frac{u}{c^2}E'_z + B'_y). \end{aligned}$$

These expressions give us the most general transformations of E'_z and B'_y under a Lorentz boost in the $\boldsymbol{e_x}$-direction. To find how the components E'_z and B'_x transform one may consider two boosts in the $\boldsymbol{e_y}$ direction of the ideal capacitor in frame S. Commencing with the field in equation (8.2) but now, recycling our notation, S' will be a frame moving with velocity $\boldsymbol{v} = v\,\boldsymbol{e_y}$ relative to S so that two surface currents on either plate are[2] $\boldsymbol{K}_{z=0} = -\sigma' v\,\boldsymbol{e_y}$ and $\boldsymbol{K}_{z=d} = \sigma' v\,\boldsymbol{e_y}$. These produce a non-zero magnetic field between the plates of $\boldsymbol{B'} = -\mu_0 K\,\boldsymbol{e_x} = -\mu_0\sigma' v\,\boldsymbol{e_x}$. In frame S' the non-zero components of the electric and magnetic field are

$$E'_z = \frac{\sigma'}{\epsilon_0} = \gamma(v)\frac{\sigma}{\epsilon_0} \qquad \text{and} \qquad B'_x = -\mu_0 v\gamma(v)\sigma.$$

As before we now view this system from another frame S'' which is now moving with speed $\boldsymbol{u} = u\,\boldsymbol{e_y}$ relative to S'. In S'' we have

$$E''_z = \frac{\sigma''}{\epsilon_0} = \gamma(w)\frac{\sigma}{\epsilon_0} \qquad \text{and} \qquad B''_x = -\mu_0 w\sigma'' = -\mu_0 w\gamma(w)\sigma.$$

where $w = u \oplus_R v = \frac{u+v}{1+\frac{uv}{c^2}}$. Compared to the system boosted in the $\boldsymbol{e_x}$ direction we have a non-zero B''_x component (and a sign difference) but otherwise the result is similar. We may rewrite E''_z and B''_x in terms of E'_z and B'_x to obtain:

$$\begin{aligned} E''_z &= \gamma(u)\big(E'_z - uB'_x\big) \qquad \text{and} \\ B''_x &= \gamma(u)\big(-\frac{u}{c^2}E'_z + B'_x\big). \end{aligned} \tag{8.3}$$

These are the Lorentz transformations of E'_z and B'_x under a Lorentz boost by $\boldsymbol{u} = u\,\boldsymbol{e_y}$.

The Lorentz boosts of the system in the z-direction contract the distance between the plates but do not change the electric field. For a frame S' moving with $\boldsymbol{v} = v\,\boldsymbol{e_z}$ with respect to S there is no magnetic field observed in S'. Now a second boost of the ideal capacitor by $\boldsymbol{u} = u\,\boldsymbol{e_z}$ to a frame S'' reveals an unchanged electric field and still no magnetic field. We have,

$$E''_z = E'_z.$$

[2]Follow the positively charged plate to deduce the direction of $\boldsymbol{K}$ as viewed in S' - recall that the direction of the current is the direction of flow of *positive* charge.

The usefulness of considering the transformation of the ideal capacitor was that it gave a system in which there were uniform electric and magnetic fields which could be transformed. However in the last case the boost of the capacitor in the direction perpendicular to the plates did not produce a magnetic field: we do not yet know how B'_z transforms under a Lorentz boost in the $\boldsymbol{e_z}$ direction. We have exhausted all the boosts of the ideal capacitor in three dimensions, so we must think of a different simple setting in which we may boost a magnetic field in a direction parallel to the field.

The infinite solenoid centred about the z-axis, considered in example 5.5, has a non-zero magnetic field within the solenoid given by $\boldsymbol{B} = \mu_0 nI\,\boldsymbol{e_z}$ where I is the current flowing (in the $\boldsymbol{e_\theta}$ direction) in each of n turns of wire per unit length making up the solenoid. Now suppose that $n = \frac{n_L}{L}$ where n_L is the number of turns of wire in a length L along the solenoid (i.e. in the $\boldsymbol{e_z}$-direction) in frame S, where the solenoid is at rest. In a second frame S' moving at constant velocity $\boldsymbol{u} = u\,\boldsymbol{e_z}$ compared to S, the solenoid is moving with velocity $-u\,\boldsymbol{e_z}$ and the distance measured in the z-direction is observed to be contracted so $L' = \frac{1}{\gamma(u)}L$. Hence the number of loops of wire in S' has increased to $n' = \gamma(u)n$. At this stage we suspect that the magnetic field in S' may increase, however we must also consider the effect of the Lorentz transformation on the current. Since I measures the rate of flow of charge, then, due to time dilation, it is observed in S' that the charges move more slowly, so less charge passes a specified point on each loop of wire in a unit of time (the definition of the current), so the current is reduced to $I' = \frac{1}{\gamma(u)}I$ in frame S'. Both these changes cancel out in the expression for the magnetic field in S':

$$\boldsymbol{B}' = \mu_0 n' I'\,\boldsymbol{e_z} = \mu_0(\gamma(u)n)\frac{I}{\gamma(u)}\,\boldsymbol{e_z} = \mu_0 nI\,\boldsymbol{e_z} = \boldsymbol{B}.$$

Hence B_z is not transformed by a Lorentz boost in the $\boldsymbol{e_z}$ direction.

What we have seen is that a Lorentz boost transforms the components of the electric and magnetic field which are perpendicular to the direction of the boost, but leaves the components parallel to the boost direction unchanged. From the results above, we can see how pairs of components of the electric and magnetic field, perpendicular to the Lorentz transformation transform. We can see directly how E'_z and B'_y are transformed into each other under a Lorentz boost with velocity in the $\boldsymbol{e_x}$ direction. The other pair of perpendicular field components are E'_y and B'_z, which we wish to boost in the $\boldsymbol{e_x}$ direction. We see the boost by $u\,\boldsymbol{e_y}$ of (E'_z, B'_x) in (8.3), but this is equivalent to boosting (E'_y, B'_z) by $u\,\boldsymbol{e_x}$, we simply replace $E'_z \to E'_y$, and $B'_x \to B'_z$ to find: $E''_y = \gamma(u)\big(E'_y - uB'_z\big)$ and $B''_z = \gamma(u)\big(-\frac{u}{c^2}E'_y + B'_z\big)$ for a boost by $\boldsymbol{u} = u\,\boldsymbol{e_x}$.

We may now collect together our observations and write the Lorentz transformation of $\boldsymbol{E}$ and $\boldsymbol{B}$ due to a Lorentz boost in the x-direction. Let S and S' be two inertial frames, where

S' has velocity $\boldsymbol{v} = v\,\boldsymbol{e_x}$ with respect to S, then

$$\begin{aligned} E'_x &= E_x & B'_x &= B_x \\ E'_y &= \gamma(v)(E_y - vB_z) & B'_y &= \gamma(v)(\frac{v}{c^2}E_z + B_y) \\ E'_z &= \gamma(v)(E_z + vB_y) & B'_z &= \gamma(v)(-\frac{v}{c^2}E_y + B_z). \end{aligned} \tag{8.4}$$

These are complicated transformations - the electric and magnetic fields do not transform as the spatial part of a Lorentz vector.

In certain circumstances the transformations may be usefully simplified. Notice that if $\boldsymbol{B} = \boldsymbol{0}$ in S, then, in S',

$$\boldsymbol{E'} = \begin{pmatrix} E_x \\ \gamma(v)E_y \\ \gamma(v)E_z \end{pmatrix}$$

$$\boldsymbol{B'} = \gamma(v)\frac{v}{c^2}\begin{pmatrix} 0 \\ E_z \\ -E_y \end{pmatrix} = \frac{v}{c^2}\begin{pmatrix} 0 \\ E'_z \\ -E'_y \end{pmatrix} = -\frac{1}{c^2}(\boldsymbol{v} \times \boldsymbol{E'})$$

where we recall that $\boldsymbol{v} = v\,\boldsymbol{e_x}$. While if $\boldsymbol{E} = \boldsymbol{0}$ in S then, in S',

$$\boldsymbol{B'} = \begin{pmatrix} B_x \\ \gamma(v)B_y \\ \gamma(v)B_z \end{pmatrix}$$

$$\boldsymbol{E'} = \gamma(v)v\begin{pmatrix} 0 \\ -B_z \\ B_y \end{pmatrix} = v\begin{pmatrix} 0 \\ -B'_z \\ B'_y \end{pmatrix} = \boldsymbol{v} \times \boldsymbol{B'}.$$

From the above we conclude that if there is a frame in which either $\boldsymbol{E}$ or $\boldsymbol{B}$ vanish then in other inertial reference frames they are related as shown above.

Example 8.1. *Use the Lorentz transformations to find the electric field of a point charge q which is moving with velocity $\boldsymbol{v} = v\,\boldsymbol{e_x}$.*

The rest-frame of the charge is S' in the Lorentz transformations above. Let us locate the charge q at the origin, then in S' we have

$$\boldsymbol{E'} = \frac{1}{4\pi\epsilon_0}\frac{q}{r'^2}\,\boldsymbol{e_r}' = \frac{1}{4\pi\epsilon_0}\Big(\frac{q}{(x'^2 + y'^2 + z'^2)^{\frac{3}{2}}}\Big)\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}$$

and

$$\boldsymbol{B'} = \mathbf{0}.$$

To find the electric and magnetic fields in S we may consider the Lorentz transformation due to a boost by $-v\,\boldsymbol{e_x}$ i.e.

$$\begin{aligned} E_x &= E'_x & B_x &= B'_x \\ E_y &= \gamma(v)(E'_y + vB'_z) & B_y &= \gamma(v)(-\frac{v}{c^2}E'_z + B'_y) \\ E_z &= \gamma(v)(E'_z - vB'_y) & B_z &= \gamma(v)(\frac{v}{c^2}E'_y + B'_z). \end{aligned}$$

We obtain

$$\boldsymbol{E} = \begin{pmatrix} E'_x \\ \gamma(v)E'_y \\ \gamma(v)E'_z \end{pmatrix} = \frac{1}{4\pi\epsilon_0}\left(\frac{q}{(x'^2 + y'^2 + z'^2)^{\frac{3}{2}}}\right)\begin{pmatrix} x' \\ \gamma(v)y' \\ \gamma(v)z' \end{pmatrix}$$

and

$$\boldsymbol{B} = \frac{1}{c^2}(\boldsymbol{v} \times \boldsymbol{E}).$$

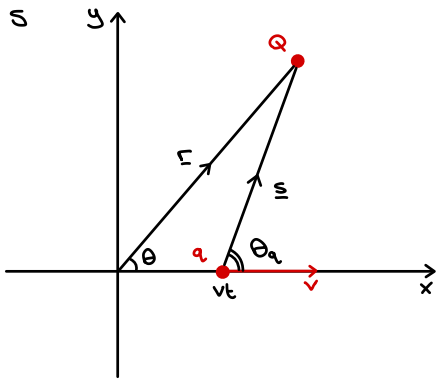
To express the fields in the coordinates of S we use the Lorentz transformation:

$$\begin{aligned} t' &= \gamma(t - \frac{v}{c^2}x) \\ x' &= \gamma(-vt + x) \\ y' &= y \\ z' &= z \end{aligned}$$

then

$$\begin{pmatrix} x' \\ \gamma(v)y' \\ \gamma(v)z' \end{pmatrix} = \gamma(v)\begin{pmatrix} x - vt \\ y \\ z \end{pmatrix} \equiv \gamma(v)\boldsymbol{s}$$

Notice that $\boldsymbol{s} = \boldsymbol{r} - vt\,\boldsymbol{e_x}$, and the source charge in S has x-coordinate vt, so that it remains at the origin in S'. Hence $\boldsymbol{s} = \boldsymbol{r} - \boldsymbol{r}_q$ where $\boldsymbol{r}_q = vt\,\boldsymbol{e_x}$ is the location of the source charge. For simplicity we will find the fields for a test charge Q located in the plane $z = 0$ depicted below:



If we adopt coordinates s and θ_q given by

$$\boldsymbol{s} = \begin{pmatrix} s\cos\theta_q \\ s\sin\theta_q \\ 0 \end{pmatrix} = \begin{pmatrix} x - vt \\ y \\ 0 \end{pmatrix}$$

then we find $x'^2 + y'^2 + z'^2 = \gamma(v)^2(x - vt)^2 + y^2 = \gamma(v)^2 s^2 \cos^2\theta_q + s^2 \sin^2\theta_q$ *and so we have,*

$$\begin{aligned} \boldsymbol{E} &= \frac{1}{4\pi\epsilon_0}\Big(\frac{\gamma(v)q}{(\gamma(v)^2 s^2\cos^2\theta_q + s^2\sin^2\theta_q)^{\frac{3}{2}}}\Big)\boldsymbol{s} \\ &= \frac{1}{4\pi\epsilon_0}\Big(\frac{\frac{1}{\gamma(v)^2}q}{(\cos^2\theta_q + \frac{1}{\gamma(v)^2}\sin^2\theta_q)^{\frac{3}{2}}}\Big)\frac{1}{s^2}\,\boldsymbol{e_s} \\ &= \frac{1}{4\pi\epsilon_0}\frac{\big(1 - \frac{v^2}{c^2}\big)}{(1 - \frac{v^2}{c^2}\sin^2\theta_q)^{\frac{3}{2}}}\frac{q}{s^2}\,\boldsymbol{e_s} \end{aligned}$$

This is an interesting result, as the electric field points from the present location of the charge and not the location of the charge at the retarded time. What is most surprising is that this same result can be derived using retarded potentials, rather than Lorentz transformations, where the information from the retarded time position of the source charge combines to give the electric field above, one that radiates away from the present location of the moving charge. The field is not radial about the location of the charge, but notice when $\theta_q \to 0$ or $\theta_q \to \pi$, i.e. in front of or behind the moving charge then the field is $\frac{1}{\gamma(v)^2}$ times the field of the stationary charge, i.e. it is weaker in front of and behind the moving charge along its direction of motion. While when $\theta_q \to \frac{\pi}{2}$ or $\theta_q \to \frac{3\pi}{2}$ then the field is $\gamma(v)$ times that of the stationary charge, it is enhanced. The field is strongest in the plane perpendicular to the motion containing the charge. The field lines bunch up towards a disc about the charge and perpendicular to the motion, an electromagnetic pancake in which the electric field is stronger than it would be if the charge were stationary. The electric and magnetic fields for a point charge moving with constant velocity were found by Oliver Heaviside in 1888, long before the advent of relativity.

8.3 A Lorentz Tensor for Electromagnetism

In equation (8.4) we see the transformation of six quantities: $E_x, E_y, E_z, B_x, B_y, B_z$ into each other under a Lorentz transformation between frames S and S' where S' has relative velocity $\boldsymbol{v} = v\,\boldsymbol{e_x}$ compared to S. A Lorentz vector has four components and a Lorentz transformation acts linearly on those four components to mix those components up. That there are six components of the electromagnetic fields which transform into each other under a Lorentz transformation means that they do not transform as a Lorentz vector. However they must transform as some

Lorentz tensor, if the laws of electromagnetism are to be preserved between inertial frames. The question is: what kind of tensor?

In section 7.4.1 we defined Lorentz tensors and showed how they transform under a Lorentz transformation Λ. One of the defining properties of a particular tensor is how many independent components it has. A Lorentz vector, in index notation, S^μ has four components as $\mu \in \{0, 1, 2, 3\}$ and may take any of four values. Consider a Lorentz tensor $T^{\mu_1\mu_2}$, or $T_{\mu_1\mu_2}$, both rank-2 tensors have 16 components, $T^{\mu_1\mu_2\mu_3}$ or $T_{\mu_1\mu_2\mu_3}$ each have 64 components (there are four labels for each μ_i index). We seek a Lorentz tensor that has 6 independent components and for this we will need to consider tensors with some additional properties. To see what we mean by extra properties first consider a four-by-four matrix, it has 16 independent components (like the $T^{\mu_1\mu_2}$ or $T_{\mu_1\mu_2}$). However we regularly consider sub-classes of four-by-four matrices, e.g. the traceless matrices (having 15 independent components, the trace condition removes one degree of freedom), or the symmetric matrices (having 10 independent degrees of freedom), or the antisymmetric matrices (having 6 independent degrees of freedom). This leads us to the question: could we find an antisymmetric rank-two tensor which contains all the components of $\boldsymbol{E}$ and $\boldsymbol{B}$, whose Lorentz transformation encodes the Lorentz transformations of (8.4)? The answer will be in the affirmative.

Suppose we have a tensor $T^{\mu_1\mu_2}$ then we may construct the symmetric part as

$$T^{(\mu_1\mu_2)} = \frac{1}{2}(T^{\mu_1\mu_2} + T^{\mu_2\mu_1})$$

and the antisymmetric part as

$$T^{[\mu_1\mu_2]} = \frac{1}{2}(T^{\mu_1\mu_2} - T^{\mu_2\mu_1}).$$

Notice the use of rounded brackets to denote a symmetrisation of the two indices and the square bracket to denote an anti-symmetrisation. By interchanging μ_1 and μ_2 you can follow the terms on the right-hand-side to see that $T^{[\mu_1\mu_2]}$ picks up an overall minus sign, while $T^{(\mu_1\mu_2)}$ does not. This is the same idea that one is used to in working with the transpose of a matrix (which interchanges the two index labels): an antisymmetric matrix picks up a sign under a transpose, while a symmetric matrix is left unchanged. We may also think about the trace of a rank 2 tensor, in this case the notion is a little bit different but still similar one may contract the indices with the Minkowski metric to forma scalar which is called the trace:

$$\eta_{\mu_1\mu_2}T^{\mu_1\mu_2} = -T^{00} + T^{11} + T^{22} + T^{33} \equiv T.$$

Each of these ideas may be generalised to tensors of different rank.

An antisymmetric rank 2 tensor $T^{[\mu_1\mu_2]}$ has 6 independent components:

$$T^{[\mu_1\mu_2]} = \begin{pmatrix} 0 & t^{01} & t^{02} & t^{03} \\ -t^{01} & 0 & t^{12} & t^{13} \\ -t^{02} & -t^{12} & 0 & t^{23} \\ -t^{03} & -t^{13} & -t^{23} & 0 \end{pmatrix}.$$

How do they transform under the Lorentz boost in the x-direction? The matrix Λ (used to express quantities in S' in terms of those in S, which is what we need to compare with equation (8.4)) is

$$\Lambda = \begin{pmatrix} \gamma & -\gamma\frac{v}{c} & 0 & 0 \\ -\gamma\frac{v}{c} & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where $\gamma = \gamma(v)$. A contravariant rank-2 tensor transforms under $\boldsymbol{S} \to \Lambda\boldsymbol{S}$ as

$$T^{[\mu_1\mu_2]} \to \Lambda^{\mu_1}{}_{\nu_1}\Lambda^{\mu_2}{}_{\nu_2}T^{[\nu_1\nu_2]}$$

Let us consider each non-zero component. For $\mu_1 = 0, \mu_2 = 1$ we have

$$\begin{aligned} t'^{01} &= \Lambda^0{}_{\nu_1}\Lambda^1{}_{\nu_2}T^{[\nu_1\nu_2]} \\ &= \Lambda^0{}_{\nu_1}\Lambda^1{}_0T^{[\nu_1 0]} + \Lambda^0{}_{\nu_1}\Lambda^1{}_1T^{[\nu_1 1]} + \Lambda^0{}_{\nu_1}\Lambda^1{}_2T^{[\nu_1 2]} + \Lambda^0{}_{\nu_1}\Lambda^1{}_3T^{[\nu_1 3]} \\ &= \Lambda^0{}_{\nu_1}\Lambda^1{}_0T^{[\nu_1 0]} + \Lambda^0{}_{\nu_1}\Lambda^1{}_1T^{[\nu_1 1]} \\ &= \Lambda^0{}_0\Lambda^1{}_0T^{[00]} + \Lambda^0{}_0\Lambda^1{}_1T^{[01]} + \Lambda^0{}_1\Lambda^1{}_0T^{[10]} + \Lambda^0{}_1\Lambda^1{}_1T^{[11]} \\ &= \Lambda^0{}_0\Lambda^1{}_1t^{01} - \Lambda^0{}_1\Lambda^1{}_0t^{01} \\ &= \gamma^2t^{01} - \gamma^2\frac{v^2}{c^2}t^{01} \\ &= t^{01} \end{aligned}$$

where in the second line we have used the fact that $\Lambda^1{}_2 = 0$ and $\Lambda^1{}_3 = 0$, and in writing the third line that $\Lambda^0{}_2 = 0$ and $\Lambda^0{}_3 = 0$, in the fourth line that $T^{[00]} = 0$ and $T^{[11]} = 0$. The t^{01} component is not transformed. Now let us consider the other components, we will repeat the

same steps above (but this time we omit the detailed working)

$$
\begin{aligned}
t'^{02} &= \Lambda^0{}_{\nu_1}\Lambda^2{}_{\nu_2}T^{[\nu_1\nu_2]} = \Lambda^0{}_0\Lambda^2{}_2T^{[02]} + \Lambda^0{}_1\Lambda^2{}_2T^{[12]} = \gamma(t^{02} - \frac{v}{c}t^{12}) \\
t'^{03} &= \Lambda^0{}_{\nu_1}\Lambda^3{}_{\nu_2}T^{[\nu_1\nu_2]} = \Lambda^0{}_0\Lambda^3{}_3T^{[03]} + \Lambda^0{}_1\Lambda^3{}_3T^{[13]} = \gamma(t^{03} - \frac{v}{c}t^{13}) \\
t'^{12} &= \Lambda^1{}_{\nu_1}\Lambda^2{}_{\nu_2}T^{[\nu_1\nu_2]} = \Lambda^1{}_0\Lambda^2{}_2T^{[02]} + \Lambda^1{}_1\Lambda^2{}_2T^{[12]} = \gamma(-\frac{v}{c}t^{02} + t^{12}) \\
t'^{13} &= \Lambda^1{}_{\nu_1}\Lambda^3{}_{\nu_2}T^{[\nu_1\nu_2]} = \Lambda^1{}_0\Lambda^3{}_3T^{[03]} + \Lambda^1{}_1\Lambda^3{}_3T^{[13]} = \gamma(-\frac{v}{c}t^{03} + t^{13}) \\
t'^{23} &= \Lambda^2{}_{\nu_1}\Lambda^3{}_{\nu_2}T^{[\nu_1\nu_2]} = \Lambda^2{}_2\Lambda^3{}_3T^{[23]} = t^{23}.
\end{aligned}
$$

By careful comparison with (8.4) we can match[3] the six components $t^{\mu_1\mu_2}$ with the components of $\boldsymbol{E}$ and $\boldsymbol{B}$:

$$
\begin{aligned}
t^{01} = \frac{E_x}{c} \qquad t^{02} = \frac{E_y}{c} \qquad t^{03} = \frac{E_z}{c} \\
t^{12} = B_z \qquad t^{13} = -B_y \\
t^{23} = B_x.
\end{aligned}
$$

Hence the components of the electromagnetic fields transform as part of a rank 2 tensor, called the electromagnetic field tensor[4]:

$$
F^{\mu\nu} = \begin{pmatrix} 0 & \frac{E_x}{c} & \frac{E_y}{c} & \frac{E_z}{c} \\ -\frac{E_x}{c} & 0 & B_z & -B_y \\ -\frac{E_y}{c} & -B_z & 0 & B_x \\ -\frac{E_z}{c} & B_y & -B_x & 0 \end{pmatrix}
$$

and we may construct the covariant rank 2 antisymmetric tensor by lowering the indices of the components above using the Minkowski metric: $F_{\mu\nu} = \eta_{\mu\lambda}\eta_{\nu\kappa}F^{\lambda\kappa}$. We find,

$$
F_{\mu\nu} = \begin{pmatrix} 0 & -\frac{E_x}{c} & -\frac{E_y}{c} & -\frac{E_z}{c} \\ \frac{E_x}{c} & 0 & B_z & -B_y \\ \frac{E_y}{c} & -B_z & 0 & B_x \\ \frac{E_z}{c} & B_y & -B_x & 0 \end{pmatrix}.
$$

We note that we can construct the Lorentz invariant

$$
F_{\mu\nu}F^{\mu\nu} = 2(\boldsymbol{B}^2 - \frac{1}{c^2}\boldsymbol{E}^2).
$$

In the electromagnetic field tensor we have unified the electric and magnetic fields in a single tensor. It remains to see how Maxwell's equations may be fully reformulated in this covariant notation.

[3] We can actually do this in two different ways, the alternative choice is to replace $\frac{\boldsymbol{E}}{c} \to \boldsymbol{B}$ and $\boldsymbol{B} \to -\frac{\boldsymbol{E}}{c}$ in the expressions that follow. This second choice constructs an interesting tensor called the dual field strength tensor $\star F_{\mu_1\mu_2} = \frac{1}{2}\epsilon_{\mu_1\mu_2\mu_3\mu_4}F^{\mu_3\mu_4}$ which is very useful in the fullest Lorentz covariant formulation of electromagnetism.

[4] Or sometimes simply the Maxwell field strength tensor, or even just the field strength.

8.4 Maxwell's Equations in Tensor Notation.

To express Maxwell's equations in a Lorentz covariant way, we must express the source charge density ρ and the source current density in a Lorentz covariant way. There are four terms, and so we may hope they can be combined into a Lorentz vector. We must investigate how the sources transform under a Lorentz transformation.

Consider a volume V containing uniform charge Q at moving with constant velocity $\boldsymbol{v}$ in a frame S then the rest-frame charge density $\rho_0 = \frac{Q}{V_0}$ where V_0 is the volume of the charge as measured in its rest-frame. As viewed from S the volume V_0 is contracted along the direction of the motion, so that in S the charge is contained in a volume $V = \frac{1}{\gamma}V_0$ where $\gamma = \gamma(v)$, hence the charge density in S is

$$\rho = \frac{Q}{V} = \gamma\frac{Q}{V_0} = \gamma\rho_0.$$

There is a current volume density associated with the volume of moving charge given by $\boldsymbol{J} = \rho\boldsymbol{v} = \rho_0\gamma\boldsymbol{v}$. We may combine these terms together in the form $(c\rho, \boldsymbol{J})$ to find the current four-vector, or four-current:

$$\boldsymbol{J} = \begin{pmatrix} c\rho \\ J_x \\ J_y \\ J_z \end{pmatrix} = \rho_0\gamma(v)\begin{pmatrix} c \\ v_x \\ v_y \\ v_z \end{pmatrix} = \rho_0\boldsymbol{V}$$

since ρ_0 is a constant then the four-current $\boldsymbol{J}$ is proportional to the four-velocity and so transforms as a contravariant four-vector. The continuity equation $\frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} + \frac{\partial J_z}{\partial z} = -\frac{\partial \rho}{\partial t}$ may be rewritten as

$$0 = \frac{\partial \rho}{\partial t} + \frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} + \frac{\partial J_z}{\partial z} \equiv \partial_\mu J^\mu$$

where

$$\partial_\mu = \begin{pmatrix} \frac{1}{c}\frac{\partial}{\partial t} \\ \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix}$$

giving a simple four-vector version of the continuity equation $\partial_\mu J^\mu = 0$. The covariant differential operator ∂_μ we introduced has a contravariant counterpart:

$$\partial^\mu = \begin{pmatrix} -\frac{1}{c}\frac{\partial}{\partial t} \\ \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix}$$

so that

$$\partial^2 = \partial_\mu \partial^\mu = -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} = \Box.$$

Two of Maxwell's equations may be written as

$$\partial_\nu F^{\mu\nu} = \mu_0 J^\mu.$$

When $\mu = 0$, $\partial_\nu F^{\mu\nu} = \mu_0 J^\mu$ becomes

$$\begin{aligned}
\partial_\nu F^{0\nu} &= \frac{\partial}{\partial x}(F^{01}) + \frac{\partial}{\partial y}(F^{02}) + \frac{\partial}{\partial z}(F^{03}) \\
&= \frac{1}{c}\frac{\partial E_x}{\partial x} + \frac{1}{c}\frac{\partial E_y}{\partial y} + \frac{1}{c}\frac{\partial E_z}{\partial z} \\
&= \frac{1}{c}\boldsymbol{\nabla}\cdot\boldsymbol{E} \\
&= \mu_0 J^0 \\
&= \mu_0 c\rho
\end{aligned}$$

which becomes (using $c^2 = \frac{1}{\epsilon_0\mu_0}$) $\boldsymbol{\nabla}\cdot\boldsymbol{E} = \frac{\rho}{\epsilon_0}$. For $\mu = i \in \{1,2,3\}$ the equation gives

$$\begin{aligned}
\partial_\nu F^{i\nu} &= \frac{1}{c}\frac{\partial}{\partial t}(F^{i0}) + \frac{\partial}{\partial x}(F^{i1}) + \frac{\partial}{\partial y}(F^{i2}) + \frac{\partial}{\partial z}(F^{i3}) \\
&= \mu_0 J^i.
\end{aligned}$$

Now we can rewrite the components of the field tensor as $F^{i0} = -\frac{1}{c}E^i$ and the component $F^{ij} = \epsilon^{ijk}B_k$ (for $i,j,k \in \{1,2,3\}$ where $x^1 = x$, $x^2 = y$ and $x^3 = z$), hence we have

$$\begin{aligned}
\partial_\nu F^{i\nu} &= -\frac{1}{c^2}\frac{\partial E^i}{\partial t} + \partial_j(\epsilon^{ijk}B_k) \\
&= \mu_0 J^i.
\end{aligned}$$

which in vector notation becomes

$$-\frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t} + \boldsymbol{\nabla}\times\boldsymbol{B} = \mu_0\boldsymbol{J}.$$

The remaining pair of Maxwell's equations arise from a property of the field strength tensor, called the Bianchi identity,

$$\partial_{[\mu}F_{\nu\rho]} \equiv \frac{1}{3}(\partial_\mu F_{\nu\rho} + \partial_\nu F_{\rho\mu} + \partial_\rho F_{\mu\nu}) = 0$$

we will prove this holds when we look at the potentials as four-vectors. For now let us show that it encodes the remaining Maxwell equations: The expression is totally antisymmetric so

is only non-trivial when μ, ν, ρ take distinct values. If each of the indices is space-like then (for $i, j, k \in \{1, 2, 3\}$):

$$\begin{aligned} 0 &= \frac{1}{3}\big(\partial_i F_{jk} + \partial_j F_{ki} + \partial_k F_{ij}\big) \\ &= \frac{1}{3}\big(\partial_i(\epsilon_{jkl}B^l) + \partial_j(\epsilon_{kil}B^l) + \partial_k(\epsilon_{ijl}B^l)\big) \end{aligned}$$

we note that $F_{ij} = \epsilon_{ijk}B^k$. If we multiply this expression with ϵ^{ijk} (and sum over the repeated indices using $\epsilon^{ijk}\epsilon_{ijl} = 2\delta^k_l$ - see appendix A) we have

$$\begin{aligned} 0 &= \epsilon^{ijk}\frac{1}{3}\big(\partial_i(\epsilon_{jkl}B^l) + \partial_j(\epsilon_{kil}B^l) + \partial_k(\epsilon_{ijl}B^l)\big) \\ &= \frac{1}{3}\big(\partial_i(2\delta^i_l B^l) + \partial_j(2\delta^j_l B^l) + \partial_k(2\delta^k_l B^l)\big) \\ &= 2\,\partial_i B^i \\ &= 2\,\boldsymbol{\nabla}\cdot\boldsymbol{B}. \end{aligned}$$

Now consider the Bianchi identity when $\mu = 0$, $\nu = i$ and $\rho = j$ for $i, j \in \{1, 2, 3\}$, we note that $F_{0i} = -\frac{E_i}{c}$, and we find

$$\begin{aligned} 0 &= \frac{1}{3}\big(\partial_0 F_{ij} + \partial_i F_{j0} + \partial_j F_{0i}\big) \\ &= \frac{1}{3}\big(\partial_0(\epsilon_{ijl}B^l) + \partial_i(\frac{E_j}{c}) + \partial_j(-\frac{E_i}{c})\big). \end{aligned}$$

Multiplying this equation by ϵ^{ijk} we obtain (dropping the coefficients),

$$0 = \frac{\partial B^k}{\partial t} + \epsilon^{ijk}\partial_i E_j$$

or, in vector notation,

$$\mathbf{0} = \frac{\partial \boldsymbol{B}}{\partial t} + \boldsymbol{\nabla}\times\boldsymbol{E}$$

which gives the last of Maxwell's equations.

To see why the Bianchi identity holds we must look at the electric and magnetic potentials in Minkowski space. You will not be surprised to learn that the electric potential Φ and the magnetic potential $\boldsymbol{A}$ combine to form a four-vector. As we saw $\Box = \partial_\mu\partial^\mu$ is a Lorentz scalar hence if the inhomogeneous wave equations $\Box\Phi = -\frac{\rho}{\epsilon_0}$ and $\Box\boldsymbol{A} = -\mu_0\boldsymbol{J}$ are to remain valid they must be recast into a Lorentz tensor equation, and the simplest possibility is if the potentials combine to form the four-potential:

$$A^\mu = \begin{pmatrix} \frac{1}{c}\Phi \\ A_x \\ A_y \\ A_z \end{pmatrix}$$

then the inhomogeneous wave equations are

$$\Box A^\mu = -\mu_0 J^\mu$$

and the electromagnetic field tensor, in terms of the four-potential, is

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$$

while

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu.$$

Now we observe the origin of the Bianchi identity as

$$\begin{aligned}
\partial_{[\mu} F_{\nu\rho]} &= \frac{1}{3}(\partial_\mu F_{\nu\rho} + \partial_\nu F_{\rho\mu} + \partial_\rho F_{\mu\nu}) \\
&= \frac{1}{3}(\partial_\mu(\partial_\nu A_\rho - \partial_\rho A_\nu) + \partial_\nu(\partial_\rho A_\mu - \partial_\mu A_\rho) + \partial_\rho(\partial_\mu A_\nu - \partial_\nu A_\mu)) \\
&= \frac{1}{3}(\partial_\mu\partial_\nu A_\rho - \partial_\mu\partial_\rho A_\nu + \partial_\nu\partial_\rho A_\mu - \partial_\nu\partial_\mu A_\rho + \partial_\rho\partial_\mu A_\nu - \partial_\rho\partial_\nu A_\mu) \\
&= 0
\end{aligned}$$

since $\partial_\mu\partial_\nu = \partial_\nu\partial_\mu$.

We notice that the electromagnetic tensor is unchanged under the transformation of the four-potential $A^\mu \to A^\mu + \partial^\mu\Lambda$ as

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu \to F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu + \partial^\mu\partial^\nu\Lambda - \partial^\nu\partial^\mu\Lambda = F^{\mu\nu}$$

so that the electric and magnetic fields are unchanged by this transformation. This is the four-vector formulation of a gauge transformation. In terms of the components we recover $\frac{1}{c}\Phi \to \frac{1}{c}\Phi - \frac{1}{c}\frac{\partial}{\partial t}\Lambda$, and $\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\Lambda$ as we found in section 6.4.

Thus all of electromagnetism reduces to a pair of equations written in Lorentz covariant form:

$$\partial_\nu F^{\mu\nu} = \mu_0 J^\mu \qquad \text{and} \qquad \partial_{[\mu} F_{\nu\rho]} = 0$$

where $\partial_{[\mu} F_{\nu\rho]} = \frac{1}{3}(\partial_\mu F_{\nu\rho} + \partial_\nu F_{\rho\mu} + \partial_\rho F_{\mu\nu})$.

Appendices

A. Index Notation and the Einstein Summation Convention

Index notation is very useful to help develop and sometimes even express fundamental scalar and vector equations. We will adopt some standard conventions in this course, the leading feature being we will adopt a different meaning for an upper index and a lower index on a vector component.

Let a vector $\boldsymbol{v}$ have components v^i, as $\boldsymbol{v}$ is a column vector then the index i labels the rows of the single column in the vector. For example, suppose that $\boldsymbol{v} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$ then the components are $v^1 = a$, $v^2 = b$ and $v^3 = c$. A scalar function, just like the components of a vector, has an output which is a single number, so in index notation it does not differ and it does not have any indices. For example the scalar function f in components is f. The indices on a component are used to specify which element in an array of numbers is being referred to. Now we know that the scalar product of a vector with itself gives a scalar and that we can write the scalar product of $\boldsymbol{v}$ with itself in terms of matrix multiplication as $\boldsymbol{v}^T\boldsymbol{v}$, where $\boldsymbol{v}^T$ denotes the transpose of $\boldsymbol{v}$. As $\boldsymbol{v}^T$ is a row vector, its components will be labelled by a column number. Our convention will be to use a lower label on a vector to denote a column label, hence the i'th component of $\boldsymbol{v}^T$ is given by v_i. The fact that the index is down indicates that we are considering an object with column labels derived from the vector $\boldsymbol{v}$, and is sufficient to inform us that v_i is a component of the transpose of $\boldsymbol{v}$. That is we may write $\boldsymbol{v} \cdot \boldsymbol{v} = \boldsymbol{v}^T\boldsymbol{v} = \sum_{i=1}^{3} v_i v^i$. The important take-away is that an upper index labels a row, while a lower index labels a column.

Let us next consider matrices, which have rows and columns - so we expect a matrix component to have one upper and one lower index, i.e. $A^i{}_j$ indicates the component of the matrix A in the i'th row and j'th column. In the scalar product, matrix multiplication in components involves a sum over one upper and one lower index. How does the multiplication of a matrix A and a vector $\boldsymbol{v}$ look in components? Well, suppose that $A\boldsymbol{v} = \boldsymbol{u}$ where $\boldsymbol{u}$ is a column vector,

so that it has components u^i. Consequently we expect that $A\boldsymbol{v}$ written in components will also have only one free index (meaning it has not been summed over) labelled i, and furthermore this must be an upper index. We can write $(A\boldsymbol{v})^i = u^i$ to indicate the i'th equation implied by $A\boldsymbol{v} = \boldsymbol{u}$, but this hides how we write the multiplication of a matrix and vector in components. As matrix multiplication involves a sum in each row across the column labels (lower index) of a matrix and down the vector labelled by row indices (upper index), then we deduce that

$$(A\boldsymbol{v})^i = \sum_{j=1}^{3} A^i{}_j v^j = u^i.$$

Our conventions mean that a sum over neighbouring pair of indices, one on the left lower and an upper index to the right, labelled by the same index encodes matrix multiplication.

The Einstein summation convention adopts the convention that repeated pairs of indices (those denoted by the same letter) are summed over without having to write the summation symbol, that is $\sum_{i=1}^{3} u_i v^i$ is written as just $u_i v^i$ and the Einstein summation convention implies that the repeated i index is summed over its full range. In three dimensions, adopting the Einstein summation convention, we have

$$u_i v^i = u_1 v^1 + u_2 v^2 + u_3 v^3$$

and the matrix equation $A\boldsymbol{v} = \boldsymbol{u}$ is, in components,

$$A^i{}_j v^j = u^i.$$

What about the transposed equation $(A\boldsymbol{v})^T = \boldsymbol{v}^T A^T = \boldsymbol{u}^T$? There is some subtlety involved in deducing the expression in components for a transposed matrix, the relation between the components of a matrix A and its transpose A^T is

$$(A^T)^i{}_j = A_j{}^i$$

in this we see that the first entry on the matrix component labels the row, and the second the column - but whether the first index is up or down informs us whether we work with the components of A^T (having components $A_j{}^i$) or A (having components $A^i{}_j$). To emphasise the notational choice, consider a two-dimensional example: let $A = \boldsymbol{v}\boldsymbol{u}^T = \begin{pmatrix} v^1u_1 & v^1u_2 \\ v^2u_1 & v^2u_2 \end{pmatrix} \equiv \begin{pmatrix} A^1{}_1 & A^1{}_2 \\ A^2{}_1 & A^2{}_2 \end{pmatrix}$, where in labelling the components we note that the first index on $A^i{}_j$ is the column vector index associated to the component of $\boldsymbol{v}$ and the second (row vector) index is associated to $\boldsymbol{u}$. Now $A^T = \boldsymbol{u}\boldsymbol{v}^T = \begin{pmatrix} u^1v_1 & u^1v_2 \\ u^2v_1 & u^2v_2 \end{pmatrix} = \begin{pmatrix} v_1u^1 & v_2u^1 \\ v_1u^2 & v_2u^2 \end{pmatrix} = \begin{pmatrix} A_1{}^1 & A_2{}^1 \\ A_1{}^2 & A_2{}^2 \end{pmatrix}$ and so $(A^T)^i{}_j = A_j{}^i$.

Returning to consider $(A\boldsymbol{v})^T = \boldsymbol{v}^T A^T = \boldsymbol{u}^T$ we write in index notation $v_j (A^T)^j{}_i = u_i$, and using that $(A^T)^i{}_j = A_j{}^i$ we find $v_j A_i{}^j$. Note that this final expression is not a matrix multiplication of a row vector $\boldsymbol{v}^T$ with the matrix A, because the repeated indices are not nearest neighbours (with the left index down and the right index up) - to read off the matrix multiplication one must go back to $v_j (A^T)^j{}_i$ to determine this is actually $\boldsymbol{v}^T A^T$ in vector notation. One message to take away is that both the nature of an index (whether it is up or down) on a matrix, and its position (whether it is the first or second index) play a role in interpreting expressions in index notation.

Comment(s).

- *A repeated index which is summed over is called a dummy index or summation index, any other index is called a free index.*

- *By repeating an index in the Einstein summation convention we add an operation (the sum over that index) to that label. Consequently there are some strict rules we must follow when writing equations in components using this convention:*

  - *Repeated indices are implicitly summed over.*

  - *No index can appear more than twice in any term, otherwise we would not know how to expand the sum. E.g. consider the forbidden expression $u_j A^i{}_j v^j$, there are three indices labelled j here, so we are left wondering which pair (or even triple) of indices we should sum over, i.e. does this mean: $u_1 A^i{}_1 v^j + u_2 A^i{}_2 v^j + u_3 A^i{}_3 v^j$, $u_1 A^i{}_j v^1 + u_2 A^i{}_j v^2 + u_3 A^i{}_j v^3$, $u_j A^i{}_1 v^1 + u_j A^i{}_2 v^2 + u_j A^i{}_3 v^3$ or even $u_1 A^i{}_1 v^1 + u_2 A^i{}_2 v^2 + u_3 A^i{}_3 v^3$?*

  - *Each term must contain identical free indices.*

- *One advantage of working in components is that all worries about non-abelian multiplication of matrices, for example, are taken care of using the components. For example while $AB \neq BA$ in general, when working in components, as each component represents a number in a matrix then these entries can be moved past each other without worry, i.e. $(AB)^i{}_j = A^i{}_k B^k{}_j = B^k{}_j A^i{}_k$. Note that $(BA)^i{}_j = B^i{}_k A^k{}_j$ which is a different expression. As another example $\boldsymbol{u}^T A \boldsymbol{v} = u_i A^i{}_k v^k = A^i{}_k u_i v^k = A^i{}_k v^k u_i$.*

- *Consider $(\boldsymbol{u}^T A \boldsymbol{v})^T = (\boldsymbol{v}^T A^T \boldsymbol{u}) = v_k (A^T)^k{}_i u^i = v_k A_i{}^k u^i$. Now $u_i A^i{}_k v^k = \boldsymbol{u}^T A \boldsymbol{v}$ is a number, so it is equal to its own transpose. Therefore we have $u_i A^i{}_k v^k = v_k A_i{}^k u^i = u^i A_i{}^k v_k$ and we glimpse another consequence of our notation: pairs of up and down indices may be raised and lowered in pairs, i.e. $A^i{}_j v^j = A^{ij} v_j$. To prove this we should introduce the Euclidean metric tensor δ_{ij} whose entries are all zero apart from $\delta_{11} = 1$, $\delta_{22} = 1$ and*

$\delta_{33} = 1$ - so it is similar to an identity matrix, albeit with a different index structure[1]*. It can be used to lower indices (i.e. carry out the transpose), so that $\delta_{ij}v^j = v_i$. We may also introduce its inverse δ^{-1} whose components are δ^{ij} with all entries zero apart from $\delta^{11} = 1$, $\delta^{22} = 1$ and $\delta^{33} = 1$, and so $\delta^{ij}v_j = v^i$. Now consider again $A^i{}_j v^j = A^i{}_j \delta^{jk} v_k = A^{ik}v_k = A^{ij}v_j$, as claimed. For the example we started with $\boldsymbol{u}^T A\boldsymbol{v} = u_i A^i{}_k v^k = u^i A_{ik} v^k = u^i A_i{}^k v_k = v_k (A^T)^k{}_i u^i = \boldsymbol{v}^T A^T \boldsymbol{u}$, where we raised and lowered the pair of i indices at the second equality and then raised and lowered the pair of k indices at the third equality. Such tricks are very useful from time to time.*

- *Another pleasant consequence of the Einstein convention is that one can readily work out the index form of any new term in an equation, as its free indices must match the free indices of all other terms. Consequently one can often rapidly complete equations by index matching against other terms and then filling in the missing indices as dummy indices. The up-down index convention we have adopted makes this even simpler, as an up free index must be a free up index on all terms in an equation (and similarly for down free indices) and we may quickly check our equations to ensure that up and down indices are matched, without ever pausing to wonder what type of object we are using nor worry too much about whether we have a row index or a column index combing in the correct way for matrix multiplication. We can rely on the conventions to do a lot of the work for us.*

Why have we belaboured these conventions for index notation here in this course? One of the consequences of unifying the electric and magnetic fields is that a four-dimensional spacetime emerges as a natural setting for electromagnetism. In this spacetime the inner product will be the Minkowski inner product, rather than the Euclidean inner product, and vectors will be upgraded to four-vectors whose zeroth component[2] is the speed of light multiplied by time ct and whose second, third and fourth components are the usual position coordinates (x, y, z) in three spatial dimensions. Let us suppose that

$$\boldsymbol{s} = \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}$$

[1] A matrix would have one up and one down index, but this object has two down indices.

[2] It is helpful to label the temporal component of a four vector as the zeroth component as then the spatial components remain the first, second and third components, so they retain their labelling from the three-dimensional setting.

then the modulus squared of the four-vector $\boldsymbol{s}$, using the Minkowski inner product[3], is

$$\boldsymbol{s}\cdot\boldsymbol{s} = s_\mu s^\mu = -c^2t^2 + x^2 + y^2 + z^2.$$

It is traditional to use Greek indices such as μ, ν, ρ for the four space-time indices and Latin indices such as i, j, k for the three spatial indices. Here we see that the lowered index has a meaning which is different from taking the simple transpose: in addition to taking the transpose the component s^1 of $\boldsymbol{s}$ has changed sign, i.e. $s_0 = -s^0$, while the spatial components have the same sign $s_1 = s^1$, $s_2 = s^2$ and $s_3 = s^3$. To keep track of the signs when we raise or lower indices we make use of the Minkowski metric tensor η which we may write as a matrix

$$\eta = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

but whose component index structure is $\eta_{\mu\nu}$ (as if it is an object with two column labels, rather than a normal matrix), so that $\eta_{00} = -1$, $\eta_{11} = 1$, $\eta_{22} = 1$ and $\eta_{33} = 1$. The importance of this index structure is that it allows us to lower indices on four-vectors as $\eta_{\mu\nu}s^\mu = s_\nu$ and hence $\boldsymbol{s}\cdot\boldsymbol{s} = s_\mu s^\mu = \eta_{\mu\nu}s^\nu s^\mu = s^\nu\eta_{\mu\nu}s^\mu = s^\nu s_\nu = s_\nu s^\nu$. Equivalently we may define its inverse η^{-1} in order to raise indices, and we write the component of η^{-1} as $\eta^{\mu\nu}$. The Kronecker delta $\delta^\mu{}_\nu$ plays the role of the identity matrix in index notation and we can deduce $\eta^{\mu\nu}$ from its defining equation $\eta^{\mu\kappa}\eta_{\kappa\nu} = \delta^\mu{}_\nu$. Therefore we adopt these vector component conventions in this course because we will meet four-dimensional space-time, and it will benefit us to get used to working with this component formulation of vector calculus in preparation.

In addition to vectors and scalars we will need to introduce a index notation for $\boldsymbol{\nabla} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix}$, as the coordinates x, y and z are replaced with the components x^1, x^2 and x^3 of a vector $\boldsymbol{x}$, then it is natural to adopt the notation ∂^i for the i'th component of $\boldsymbol{\nabla}$. Therefore $(\boldsymbol{\nabla}f)^i = \partial^i f$, and $\boldsymbol{\nabla}\cdot\boldsymbol{v} = \boldsymbol{\nabla}^T\boldsymbol{v} = \partial_i v^i$. In order to work with the components of curl we introduce the totally antisymmetric Levi-Civita tensor ϵ whose components are ϵ_{ijk}. It is called a totally-antisymmetric tensor because if any pair of neighbouring indices are interchanged[4] then it changes sign, so

$$\epsilon_{ijk} = -\epsilon_{jik} = \epsilon_{jki} = -\epsilon_{kji} = \epsilon_{kij} = -\epsilon_{ikj}$$

[3]We have picked the 'mostly-plus' convention for the Minkowski inner product here, one can work equivalently with the 'mostly-minus' convention in which $\eta \to -\eta$ and each term in $\boldsymbol{s}\cdot\boldsymbol{s}$ has the opposite sign.

[4]Interchanging neighbouring indices is similar to taking the transpose of a matrix, and a totally antisymmetric matrix is one which satisfies $A^T = -A$.

Now ϵ_{ijk} in principal has $3^3 = 27$ values (for particular values of $\{i, j, k\} \in \{1, 2, 3\}$). But due to the antisymmetry of ϵ then if any indices are the same then the component is zero, as $\epsilon_{112} = -\epsilon_{112} = 0$ as only zero is equal to its own negative. Hence only six components are non-zero: those where the indices are all distinct, and these are all related to each other by

$$\epsilon_{123} = -\epsilon_{213} = \epsilon_{231} = -\epsilon_{321} = \epsilon_{312} = -\epsilon_{132}.$$

The tensor is normalised so that $\epsilon_{123} = 1$ then we find the only non-zero components are

$$\epsilon_{123} = \epsilon_{231} = \epsilon_{312} = 1 \tag{A.1}$$

$$\epsilon_{213} = \epsilon_{321} = \epsilon_{132} = -1. \tag{A.2}$$

We can now write the transpose of $\boldsymbol{\nabla} \times \boldsymbol{v}$ in components as $(\boldsymbol{\nabla} \times \boldsymbol{v})_i = \epsilon_{ijk}\partial^j v^k$, so that, for example,

$$\begin{aligned}(\boldsymbol{\nabla} \times \boldsymbol{v})_1 &= \epsilon_{1jk}\partial^j v^k \\ &= \epsilon_{12k}\partial^2 v^k + \epsilon_{13k}\partial^3 v^k \\ &= \epsilon_{123}\partial^2 v^3 + \epsilon_{132}\partial^3 v^2 \\ &= \partial^2 v^3 - \partial^3 v^2.\end{aligned}$$

We could have given the vector component of $\boldsymbol{\nabla} \times \boldsymbol{v}$ rather than its transpose, but it is notationally less convenient[5] :

$$(\boldsymbol{\nabla} \times \boldsymbol{v})^i = \epsilon^i{}_{jk}\partial^j v^k$$

where $\epsilon^i{}_{jk} = \delta^{in}\epsilon_{njk}$. When we work with the curl we will frequently work with the row vector in order to simplify notation. Notice also that in D dimensions where the component of ϵ is $\epsilon_{i_1 i_2 \ldots i_D}$ then the same construction used for the cross-product of two vectors gives $\epsilon_{i_1 i_2 \ldots i_D} u^{i_{D-1}} v^{i_D} = (\boldsymbol{u} \times \boldsymbol{v})_{i_1 i_2 \ldots i_{D-2}}$ which is a tensor with $D-2$ indices, so only when $D = 3$ does this have the single index of a (row) vector: in general, the cross product of two vectors is only related to another vector in three dimensions.

Example A.1. *Use index notation to show that* $\boldsymbol{a} \cdot (\boldsymbol{b} \times \boldsymbol{c}) = \boldsymbol{b} \cdot (\boldsymbol{c} \times \boldsymbol{a}) = \boldsymbol{c} \cdot (\boldsymbol{a} \times \boldsymbol{b})$.

The only oddity in this example is that for notational simplicity we will work with the transpose of $\boldsymbol{b} \times \boldsymbol{c}$ *with component* $(\boldsymbol{b} \times \boldsymbol{c})_i = \epsilon_{ijk} b^j c^k$ *then to take the scalar product we must*

[5]You may worry at this stage that working with the transpose of $\boldsymbol{\nabla} \times \boldsymbol{v}$ will introduce the unnecessary complication of having to remember that it is the transpose and not a component of a column vector, but no extra memory is needed, the position of the index on the component does all the work for us. All conventions are equal, but some are just better than others.

multiply it by a^i:

$$\begin{aligned}\boldsymbol{a}\cdot(\boldsymbol{b}\times\boldsymbol{c}) &= a^i\epsilon_{ijk}b^jc^k\\ &= b^j\epsilon_{ijk}c^ka^i\\ &= -b^j\epsilon_{jik}c^ka^i\\ &= b^j\epsilon_{jki}c^ka^i\\ &= \boldsymbol{b}\cdot(\boldsymbol{c}\times\boldsymbol{a})\end{aligned}$$

and equivalently

$$\begin{aligned}\boldsymbol{a}\cdot(\boldsymbol{b}\times\boldsymbol{c}) &= a^i\epsilon_{ijk}b^jc^k\\ &= c^k\epsilon_{ijk}a^ib^j\\ &= c^k\epsilon_{kij}a^ib^j\\ &= \boldsymbol{c}\cdot(\boldsymbol{a}\times\boldsymbol{b})\end{aligned}$$

We may use the δ^{-1} to raise all the indices on the Levi-Civita tensor, and work with the similar object

$$\epsilon^{nmp} = \delta^{in}\delta^{jm}\delta^{kp}\epsilon_{ijk}.$$

One consequence of introducing ϵ^{ijk} is that we now have a way of taking the cross product of $\boldsymbol{u}^T$ and $\boldsymbol{v}^T$, i.e. we can consider objects such as $\epsilon^{ijk}u_jv_k$ which is a vector component. On the other hand $\epsilon_{ilm}u^lv^m$ is a row vector component, so we may matrix multiply these two objects together to get a scalar, i.e. $\epsilon_{ilm}u^lv^m\epsilon^{ijk}u_jv_k$ should be a scalar, and certainly all indices are dummy indices but we can simplify this expression further, by noticing that when we sum the indices of two Levi-Civita tensors some simpler object must emerge. Since $\epsilon_{ilm}\epsilon^{ijk}$ is an object with four free indices: two up and two down we must hope to build the resultant object out of the $\delta^i{}_j$ tensors. This is indeed the case, but to see this it is first helpful to consider the two-dimensional case:

$$\epsilon_{ik}\epsilon^{ij} = \epsilon_{1k}\epsilon^{1j} + \epsilon_{2k}\epsilon^{2j}.$$

Now $\epsilon_{1k}\epsilon^{1j}$ is only non-zero when $j = k = 2$, in which case $\epsilon_{12}\epsilon^{12} = 1$, and the term $\epsilon_{2k}\epsilon^{2j}$ is only non-zero if $k = j = 1$ when it gives $\epsilon_{21}\epsilon^{21} = -1 \times -1 = 1$. Hence if $k = j$ then $\epsilon_{ik}\epsilon^{ij} = 1$, i.e.

$$\epsilon_{ik}\epsilon^{ij} = \delta^j_k.$$

Now let's think about the three-dimensional case of

$$\begin{aligned}\epsilon_{ijm}\epsilon^{ijk} &= \epsilon_{1jm}\epsilon^{1jk} + \epsilon_{2jm}\epsilon^{2jk} + \epsilon_{3jm}\epsilon^{3jk}\\ &= \epsilon_{12m}\epsilon^{12k} + \epsilon_{13m}\epsilon^{13k} + \epsilon_{21m}\epsilon^{21k} + +\epsilon_{23m}\epsilon^{23k} + \epsilon_{31m}\epsilon^{31k} + \epsilon_{32m}\epsilon^{32k}\\ &= 2\delta^k_m\end{aligned}$$

by the same arguments we made for the two-dimensional case. Now let's think about

$$\epsilon_{ilm}\epsilon^{ijk} = \epsilon_{1lm}\epsilon^{1jk} + \epsilon_{2lm}\epsilon^{2jk} + \epsilon_{3lm}\epsilon^{3jk}.$$

Now the term $\epsilon_{1lm}\epsilon^{1jk}$ is only non-zero if $l = 2$ and $m = 3$ or $l = 3$ and $m = 2$ and if $j = 2$ and $k = 3$ or $j = 3$ and $k = 2$. In addition $\epsilon_{1lm}\epsilon^{1jk}$ is antisymmetric in l and m and also in j and k. We may define an antisymmetric product of δ tensors which would have the same index structure as this first term. Let us define $\delta^{jk}_{lm} = \frac{1}{2}(\delta^j_l\delta^k_m - \delta^j_m\delta^k_l)$, then notice that $\delta^{jk}_{lm} = -\delta^{jk}_{ml}$ and $\delta^{jk}_{lm} = -\delta^{kj}_{lm}$. Now if $l = 2$ and $m = 3$ then $\delta^{jk}_{23} = \frac{1}{2}(\delta^j_2\delta^k_3 - \delta^j_3\delta^k_2)$ which is equal to $\frac{1}{2}$ if $j = 2$ and $k = 3$ or $-\frac{1}{2}$ if $j = 3$ and $k = 2$ and zero otherwise. That is, $\epsilon_{1lm}\epsilon^{1jk} = 2\delta^{jk}_{lm}$, and consequently

$$\epsilon_{ilm}\epsilon^{ijk} = 2\delta^{jk}_{lm} \tag{A.3}$$

For completeness, we may also consider

$$\begin{aligned}
\epsilon_{ijk}\epsilon^{ijk} &= \epsilon_{1jk}\epsilon^{1jk} + \epsilon_{2jk}\epsilon^{2jk} + \epsilon_{3jk}\epsilon^{3jk} \\
&= \epsilon_{12k}\epsilon^{12k} + \epsilon_{13k}\epsilon^{13k} + \epsilon_{21k}\epsilon^{21k} + \epsilon_{23k}\epsilon^{23k} + \epsilon_{31k}\epsilon^{31k} + \epsilon_{32k}\epsilon^{32k} \\
&= 3!
\end{aligned}$$

and

$$\epsilon_{nmp}\epsilon^{ijk} = 3!\delta^{ijk}_{nmp}$$

where $\delta^{ijk}_{nmp} = \frac{1}{3!}(\delta^i_n\delta^j_m\delta^k_p + \delta^i_m\delta^j_p\delta^k_n + \delta^i_p\delta^j_n\delta^k_m - \delta^i_m\delta^j_n\delta^k_p - \delta^i_n\delta^j_p\delta^k_m - \delta^i_p\delta^j_m\delta^k_n)$. We may check the consistency of our results by for using the final result to compute

$$\begin{aligned}
\epsilon_{imp}\epsilon^{ijk} &= 3!\delta^{ijk}_{imp} \\
&= \delta^i_i\delta^j_m\delta^k_p + \delta^i_m\delta^j_p\delta^k_i + \delta^i_p\delta^j_i\delta^k_m - \delta^i_m\delta^j_i\delta^k_p - \delta^i_i\delta^j_p\delta^k_m - \delta^i_p\delta^j_m\delta^k_i \\
&= 3\delta^j_m\delta^k_p + \delta^j_p\delta^k_m + \delta^j_p\delta^k_m - \delta^j_m\delta^k_p - 3\delta^j_p\delta^k_m - \delta^k_p\delta^j_m \\
&= \delta^j_m\delta^k_p - \delta^j_p\delta^k_m \\
&= 2\delta^{jk}_{mp}.
\end{aligned}$$

Furthermore we also find

$$\begin{aligned}
\epsilon_{ijp}\epsilon^{ijk} &= 2\delta^{jk}_{jp} \\
&= \delta^j_j\delta^k_p - \delta^j_p\delta^k_j \\
&= 3\delta^k_p - \delta^k_p \\
&= 2\delta^k_p.
\end{aligned}$$

Example A.2. *Use index notation to show that* $\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c}) = \boldsymbol{b}(\boldsymbol{a} \cdot \boldsymbol{c}) - \boldsymbol{c}(\boldsymbol{a} \cdot \boldsymbol{b})$.

$$
\begin{aligned}
(\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c}))^i &= \epsilon^{ijk} a_j (\epsilon_{knm} b^n c^m) \\
&= a_j b^n c^m \epsilon^{kij} \epsilon_{knm} \\
&= a_j b^n c^m (\delta^i_n \delta^j_m - \delta^i_m \delta^j_n) \\
&= a_j b^n c^m \delta^i_n \delta^j_m - a_j b^n c^m \delta^i_m \delta^j_n \\
&= a_j b^i c^j - a_n b^n c^i \\
&= b^i (\boldsymbol{a} \cdot \boldsymbol{c}) - c^i (\boldsymbol{a} \cdot \boldsymbol{b}).
\end{aligned}
$$

B. Scalars and Vectors

A vector is constructed in relation to the coordinate axes of a vector space. The prototype of a vector in $\mathbb{R}^3$ is $\boldsymbol{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$. All other vectors that we construct in this vector space must transform in the same way as $\boldsymbol{r}$ under a coordinate transformation. It is simple to construct a similar array of three numbers which does not transform as a vector, for example suppose we form a possible vector whose components do not refer to the coordinates in any way, such as the number of cats, dogs and squirrels in a garden. Each component of this object is a pure number, and although we might conceive that it has a direction, it would be a "direction" in cat, dog, squirrel space and not in $\mathbb{R}^3$. This is evident upon make a change of the coordinates, for example consider a rotation of the axes of $\mathbb{R}^3$ about the z-axis by θ, under this transformation the displacement vector $\boldsymbol{r}$ transforms as

$$\boldsymbol{r} \to \boldsymbol{r'} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x\cos\theta + y\sin\theta \\ -x\sin\theta + y\cos\theta \\ z \end{pmatrix}.$$

Any vector in $\mathbb{R}^3$ must transform in an identical way under this transformation[1]. On the other hand, under a change of spatial coordinates the number of cats, dogs and squirrels does not change at all, so because it does not transform in the same way as a vector should, it is not a vector.

Scalars, like vectors, may be defined by the way they transform under a change of coordinates. However scalars, unlike vectors, do not depend on the basis vectors of $\mathbb{R}^3$. A physical

[1]Under a more general transformation if $x^i \to y^i$ where x^i is a component of $\boldsymbol{r}$ and y^i a component of $\boldsymbol{r'}$ and the transformation maps $\boldsymbol{r} \to \boldsymbol{r'}$, then the components of a vector transform as $x^i \to y^i = \frac{\partial y^i}{\partial x^j} x^j$ where the Einstein summation convention is in effect. For the rotation above, $x^1 = x$, $x^2 = y$ and $x^3 = z$, while $y^1 = x^1\cos\theta + x^2\sin\theta$, $y^2 = -x^1\sin\theta + x^2\cos\theta$, and $y^3 = x^3$, so, for example, $\frac{\partial y^1}{\partial x^j}x^j = \frac{\partial y^1}{\partial x^1}x^1 + \frac{\partial y^1}{\partial x^2}x^2 + \frac{\partial y^1}{\partial x^3}x^3$ which is indeed the case.

example of a scalar field is the temperature T at any point $\boldsymbol{r} \in \mathbb{R}^3$. If we change the coordinates of the axes, for example suppose we translate the space by $\boldsymbol{a} \in \mathbb{R}^3$, i.e. $\boldsymbol{r'} = \boldsymbol{r} + \boldsymbol{a}$, then although the way we identify points in the space has changed, the underlying temperature data has not. Mathematically we could construct the temperature scalar function in the transformed coordinate system as $T'(\boldsymbol{r} + \boldsymbol{a}) = T(\boldsymbol{r})$, as the point at $\boldsymbol{r} + \boldsymbol{a}$ in the new coordinate system will have the same temperature as the point at $\boldsymbol{r}$ in the old coordinate system. Hence the scalar field T is transformed to T' where $T' = T$. Under a change of coordinates, a scalar field is unchanged.

Physical equations should not depend upon the choice of coordinate system used, hence physical equations will be coordinate independent, which means that either side of any equation must transform in the same way under a coordinate transformation. Otherwise a change of coordinates would lead to coordinate dependent terms appearing on one side of an equation and not the other. So, for example, physical equations will have the form of 'scalar = scalar' or 'vector = vector'. More complicated variations can be considered, but the important requirement is that both sides of an equation have the same tensorial type[2].

In the preliminaries of the course we have introduced the gradient of a scalar field and have stated it is a vector, but we should check that it does indeed transform as a vector. Let's first think about why it is not clear that it is a vector. Consider $\boldsymbol{\nabla} f$ its first component is a partial derivative $\frac{\partial f}{\partial x}$. Now the derivative $\frac{\partial f}{\partial x}$ transforms differently from a scalar or a vector[3]: if we rotated the x-axis the derivative of f with respect to the new axis would generally be different. However when we combine the three derivatives $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ and $\frac{\partial f}{\partial z}$ in forming $\boldsymbol{\nabla} f$, then together they do transform as a vector. Let's consider the example of the rotation by θ about the z-axis to illustrate this idea. Under the transformation we have $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$ and $z' = z$. We will neglect the z coordinate in this example, as it does not transform. Now, using the chain rule, we have

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial x'}\frac{\partial x'}{\partial x} + \frac{\partial f}{\partial y'}\frac{\partial y'}{\partial x} = \cos\theta\frac{\partial f}{\partial x'} - \sin\theta\frac{\partial f}{\partial y'}$$
$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial x'}\frac{\partial x'}{\partial y} + \frac{\partial f}{\partial y'}\frac{\partial y'}{\partial y} = \sin\theta\frac{\partial f}{\partial x'} + \cos\theta\frac{\partial f}{\partial y'}.$$

We may rearrange this[4] to obtain the components of $\boldsymbol{\nabla} f$ in the transformed coordinate system:

$$\frac{\partial f}{\partial x'} = \cos\theta\frac{\partial f}{\partial x} + \sin\theta\frac{\partial f}{\partial y}$$
$$\frac{\partial f}{\partial y'} = -\sin\theta\frac{\partial f}{\partial x} + \cos\theta\frac{\partial f}{\partial y}.$$

[2]Scalars and vectors are simple examples of tensors.

[3]While a vector transforms in a contravariant way, a derivative transforms covariantly, to understand these terms and tensors one must study a course in differential geometry.

[4]Perhaps by noting that the inverse of a rotation by θ is a rotation by $-\theta$.

That is, if we denote the transformation of the gradient by $\boldsymbol{\nabla} f \to (\boldsymbol{\nabla} f)'$ where $(\boldsymbol{\nabla} f)' = \begin{pmatrix} \frac{\partial f}{\partial x'} \\ \frac{\partial f}{\partial y'} \\ \frac{\partial f}{\partial z'} \end{pmatrix}$

then

$$\begin{pmatrix} \frac{\partial f}{\partial x'} \\ \frac{\partial f}{\partial y'} \\ \frac{\partial f}{\partial z'} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{pmatrix}.$$

Since $\boldsymbol{\nabla} f$ transforms under the rotation in the same way as $\boldsymbol{r}$, and because, in fact, it transforms under general coordinate transformations in the same way as $\boldsymbol{r}$ then it is a vector.

C. Cartesian, Cylindrical and Spherical Polar Coordinates

We will make use of several coordinate systems throughout the course. In this appendix we will outline the changes to the expressions in Cartesian coordinates when we work in cylindrical and spherical polar coordinates.

C.1 Cylindrical Coordinates

We define our cylindrical coordinates (ρ, θ, z) in relation to Cartesian coordinates by

$$\begin{aligned} x &= \rho\cos\theta \\ y &= \rho\sin\theta \\ z &= z \end{aligned} \tag{C.1}$$

Let a set of unit basis vectors of the Cartesian coordinate system be denoted $\boldsymbol{e_x} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $\boldsymbol{e_y} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ and $\boldsymbol{e_z} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. We may find $\boldsymbol{e_x}$ as it is the change in the displacement vector $\boldsymbol{r} = x\,\boldsymbol{e_x} + y\,\boldsymbol{e_y} + z\,\boldsymbol{e_z}$ when x, and only x, is varied by one unit, y and z are held constant. We may understand $\boldsymbol{e_y}$ and $\boldsymbol{e_z}$ in a similar way. In cylindrical coordinates,

$$\boldsymbol{\rho} = \begin{pmatrix} \rho\cos\theta \\ \rho\sin\theta \\ z \end{pmatrix}$$

Suppose that we transform $\rho \to \rho + \delta\rho$ (holding θ and z fixed) then

$$\delta\boldsymbol{\rho} = \delta\rho \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix}$$

hence the unit vector in the radial ρ direction is

$$\boldsymbol{e_\rho} = \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix} = (\cos\theta)\,\boldsymbol{e_x} + (\sin\theta)\,\boldsymbol{e_y}.$$

Similarly by varying θ and z (and then normalising to unit length) we find the unit vectors

$$\boldsymbol{e_\theta} = \begin{pmatrix} -\sin\theta \\ \cos\theta \\ 0 \end{pmatrix} = (-\sin\theta)\,\boldsymbol{e_x} + (\cos\theta)\,\boldsymbol{e_y}$$
$$\boldsymbol{e_z} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \boldsymbol{e_z}.$$

Consequently we have,

$$\begin{aligned} \boldsymbol{e_x} &= (\cos\theta)\,\boldsymbol{e_\rho} - (\sin\theta)\,\boldsymbol{e_\theta} \\ \boldsymbol{e_y} &= (\sin\theta)\,\boldsymbol{e_\rho} + (\cos\theta)\,\boldsymbol{e_\theta} \\ \boldsymbol{e_z} &= \boldsymbol{e_z}. \end{aligned}$$

The infinitesimal displacements, along a path where only one of each coordinate varies, are $dl_\rho = d\rho$, $dl_\theta = \rho d\theta$ and $dl_z = dz$ and the volume element is $dv = \rho d\rho d\theta dz$.

We may rearrange (C.1) to find

$$\begin{aligned} x^2 + y^2 &= \rho^2 \\ \frac{y}{x} &= \tan\theta \\ z &= z \end{aligned}$$

then we may use the chain rule to find that

$$\begin{aligned}\boldsymbol{\nabla} f &= \frac{\partial f}{\partial x}\,\boldsymbol{e_x} + \frac{\partial f}{\partial y}\,\boldsymbol{e_y} + \frac{\partial f}{\partial z}\,\boldsymbol{e_z} \\ &= (\frac{\partial f}{\partial \rho}\frac{\partial \rho}{\partial x} + \frac{\partial f}{\partial \theta}\frac{\partial \theta}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial x})((\cos\theta)\,\boldsymbol{e_\rho} - (\sin\theta)\,\boldsymbol{e_\theta}) \\ &\quad + (\frac{\partial f}{\partial \rho}\frac{\partial \rho}{\partial y} + \frac{\partial f}{\partial \theta}\frac{\partial \theta}{\partial y} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial y})((\sin\theta)\,\boldsymbol{e_\rho} + (\cos\theta)\,\boldsymbol{e_\theta}) \\ &\quad + (\frac{\partial f}{\partial \rho}\frac{\partial \rho}{\partial z} + \frac{\partial f}{\partial \theta}\frac{\partial \theta}{\partial z} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial z})\,\boldsymbol{e_z} \\ &= \frac{\partial f}{\partial \rho}\,\boldsymbol{e_\rho} + \frac{1}{\rho}\frac{\partial f}{\partial \theta}\,\boldsymbol{e_\theta} + \frac{\partial f}{\partial z}\,\boldsymbol{e_z}. \end{aligned} \tag{C.2}$$

The divergence is usually written in terms of the inner product. The infinitesimal displacements computed above give

$$dl^2 = (d\rho)^2 + \rho^2(d\theta)^2 + (dz)^2 = d\boldsymbol{l}^T d\boldsymbol{l}$$

which suggests that the infinitesimal vector $d\boldsymbol{l}$ is

$$d\boldsymbol{l} = \begin{pmatrix} d\rho \\ \rho d\theta \\ dz \end{pmatrix}.$$

Notice the scale factor which appears on the θ component: $d\theta$ is scaled by a factor of ρ to $\rho d\theta$, in order to give a distance. Scale factors such as this are commonplace in curvilinear coordinate systems, and make some expressions more complicated than one would expect. Notice that the scale factor has made an appearance in $\boldsymbol{\nabla} f$ above: if we had moved along a path where r and z are constant i.e. $dl = \rho d\theta$ then we would expect that $\boldsymbol{\nabla} f \cdot d\boldsymbol{l} = \frac{\partial f}{\partial(\rho\theta)}\rho d\theta = \frac{1}{\rho}\frac{\partial f}{\partial \theta}(\rho d\theta)$ would give the infinitesimal change in f along a path where only θ varies. Now since $(d\boldsymbol{l})_\theta = \rho d\theta$, then $(\boldsymbol{\nabla} f)_\theta = \frac{1}{\rho}\frac{\partial}{\partial \theta}$.

Before considering the divergence directly, observe that because at different points in space the basis vectors in the cylindrical system will generally point in different directions, we must be careful when taking derivatives of the basis vectors. For example we notice that

$$\frac{\partial}{\partial \theta}\,\boldsymbol{e_\rho} = \begin{pmatrix} -\sin\theta \\ \cos\theta \\ 0 \end{pmatrix} = \boldsymbol{e_\theta}$$

$$\frac{\partial}{\partial \theta}\,\boldsymbol{e_\theta} = \begin{pmatrix} -\cos\theta \\ -\sin\theta \\ 0 \end{pmatrix} = -\boldsymbol{e_\rho}$$

but all other derivatives of the basis vectors are zero. In the divergence we will encounter such terms, and so we will be cunning. We will use the observation that the vectors

$$\frac{\boldsymbol{e_\rho}}{\rho}, \quad \boldsymbol{e_\theta}, \quad \text{and} \quad \frac{\boldsymbol{e_z}}{\rho} \tag{C.3}$$

are all divergenceless. Let us check this, because it is informative about how one computes the divergence in cases where the basis vectors vary with the coordinates. The only warning is that in computing $\boldsymbol{\nabla}\cdot\boldsymbol{v}$ one must first apply the derivatives in $\boldsymbol{\nabla}$ and then after apply the dot product (since the derivative may change the basis vector, one can see why the order of operations is important here). For example,

$$\begin{aligned}\boldsymbol{\nabla}\cdot\Big(\frac{\boldsymbol{e_\rho}}{\rho}\Big) &= \boldsymbol{e_\rho}\cdot\Big(\frac{\partial}{\partial\rho}\Big(\frac{\boldsymbol{e_\rho}}{\rho}\Big)\Big) + \boldsymbol{e_\theta}\cdot\Big(\frac{1}{\rho}\frac{\partial}{\partial\theta}\Big(\frac{\boldsymbol{e_\rho}}{\rho}\Big)\Big) + \boldsymbol{e_z}\cdot\Big(\frac{\partial}{\partial z}\Big(\frac{\boldsymbol{e_\rho}}{\rho}\Big)\Big)\\ &= \boldsymbol{e_\rho}\cdot\Big(-\frac{1}{\rho^2}\,\boldsymbol{e_\rho}\Big) + \boldsymbol{e_\theta}\cdot\Big(\frac{1}{\rho^2}\,\boldsymbol{e_\theta}\Big)\\ &= 0.\end{aligned}$$

There is another reason these three vectors in equation (C.3) are divergenceless: it is because they are all of the form $\boldsymbol{\nabla}f\times\boldsymbol{\nabla}g$. Such vectors have zero divergence, this can be seen in components as

$$\partial^i(\partial^j f\partial^k g\epsilon_{ijk}) = (\partial^i\partial^j f)\epsilon_{ijk}\partial^k g + \partial^j f(\partial^i\partial^k g)\epsilon_{ijk} = 0$$

because $\partial^i\partial^j f\epsilon_{ijk} = \partial^j\partial^i f\epsilon_{ijk} = -\partial^i\partial^j f\epsilon_{ijk}$ where we have used the fact the partial derivatives commute, followed by a relabelling of the indices. A similar computation shows that the second term above is also zero. To convince ourselves that the vectors in equation (C.3) are of this form,, we first notice (by direct computation) that

$$\boldsymbol{e_\rho}\times\boldsymbol{e_\theta} = \boldsymbol{e_z}, \quad \boldsymbol{e_\theta}\times\boldsymbol{e_z} = \boldsymbol{e_\rho}, \quad \text{and} \quad \boldsymbol{e_z}\times\boldsymbol{e_\rho} = \boldsymbol{e_\theta}.$$

Then $\boldsymbol{\nabla}\rho = \boldsymbol{e_\rho}$, $\boldsymbol{\nabla}\theta = \frac{1}{\rho}\boldsymbol{e_\theta}$, and $\boldsymbol{\nabla}z = \boldsymbol{e_z}$, hence

$$\begin{aligned}\boldsymbol{\nabla}\rho\times\boldsymbol{\nabla}\theta &= \frac{\boldsymbol{e_\rho}\times\boldsymbol{e_\theta}}{\rho} = \frac{\boldsymbol{e_z}}{\rho}\\ \boldsymbol{\nabla}\theta\times\boldsymbol{\nabla}z &= \frac{\boldsymbol{e_\theta}\times\boldsymbol{e_z}}{\rho} = \frac{\boldsymbol{e_\rho}}{\rho}\\ \boldsymbol{\nabla}z\times\boldsymbol{\nabla}\rho &= \boldsymbol{e_z}\times\boldsymbol{e_\rho} = \boldsymbol{e_\theta}.\end{aligned}$$

So each of the above vectors is divergenceless.

Now we may use this observation to compute $\boldsymbol{\nabla}\cdot\boldsymbol{v}$ where $\boldsymbol{v} = v_\rho\,\boldsymbol{e_\rho} + v_\theta\,\boldsymbol{e_\theta} + v_z\,\boldsymbol{e_z}$ by rewriting $\boldsymbol{v}$ as

$$\boldsymbol{v} = (\rho v_\rho)\frac{\boldsymbol{e_\rho}}{\rho} + v_\theta\,\boldsymbol{e_\theta} + (\rho v_z)\frac{\boldsymbol{e_z}}{\rho}$$

then using the product rule for the divergence: $\boldsymbol{\nabla}\cdot(f\boldsymbol{u}) = f\boldsymbol{\nabla}\cdot\boldsymbol{u} + (\boldsymbol{\nabla}f)\cdot\boldsymbol{u}$ where f is a scalar function and $\boldsymbol{u}$ is a vector function, we obtain,

$$\boldsymbol{\nabla}\cdot\boldsymbol{v} = (\boldsymbol{\nabla}(\rho v_\rho))\cdot\frac{\boldsymbol{e_\rho}}{\rho} + (\boldsymbol{\nabla}v_\theta)\cdot\boldsymbol{e_\theta} + (\boldsymbol{\nabla}(\rho v_z))\cdot\frac{\boldsymbol{e_z}}{\rho}$$

having eliminated the divergenceless terms from the product rule. The dot product with the basis vectors now picks out the relevant component of the various gradient functions, i.e.

$$\begin{aligned}\boldsymbol{\nabla}\cdot\boldsymbol{v} &= (\boldsymbol{\nabla}(\rho v_\rho))_\rho\frac{1}{\rho} + (\boldsymbol{\nabla}v_\theta)_\theta + (\boldsymbol{\nabla}(\rho v_z))_z\frac{1}{\rho}\\ &= \frac{1}{\rho}\frac{\partial}{\partial\rho}(\rho v_\rho) + \frac{1}{\rho}\frac{\partial}{\partial\theta}(v_\theta) + \frac{1}{\rho}\frac{\partial}{\partial z}(\rho v_z)\\ &= \frac{1}{\rho}\frac{\partial}{\partial\rho}(\rho v_\rho) + \frac{1}{\rho}\frac{\partial}{\partial\theta}(v_\theta) + \frac{\partial}{\partial z}(v_z).\end{aligned}$$

To construct the curl, we will use the fact that $\boldsymbol{\nabla}\times\boldsymbol{\nabla}f = \mathbf{0}$ for any scalar function f. In particular when $f = \rho$, θ or ϕ we have

$$\boldsymbol{\nabla}\times\boldsymbol{e_\rho} = 0, \quad \boldsymbol{\nabla}\times(\frac{1}{\rho}\boldsymbol{e_\theta}) = 0, \quad \text{and} \quad \boldsymbol{\nabla}\times\boldsymbol{e_z} = \mathbf{0}.$$

Now we may adopt the same strategy we used to find the divergence of $\boldsymbol{v}$, namely we write the vector as

$$\boldsymbol{v} = v_\rho\,\boldsymbol{e_\rho} + \rho v_\theta\Big(\frac{1}{\rho}\,\boldsymbol{e_\theta}\Big) + v_z\,\boldsymbol{e_z}$$

to make explicit the part of each term which has curl zero and then apply the product rule $\boldsymbol{\nabla}\times(f\boldsymbol{u}) = \boldsymbol{\nabla}f\times\boldsymbol{u} + f(\boldsymbol{\nabla}\times\boldsymbol{u})$ to $\boldsymbol{v}$ giving

$$\boldsymbol{\nabla}\times\boldsymbol{v} = (\boldsymbol{\nabla}v_\rho)\times\boldsymbol{e_\rho} + (\boldsymbol{\nabla}(\rho v_\theta))\times(\frac{1}{\rho}\boldsymbol{e_\theta}) + (\boldsymbol{\nabla}v_z)\times\boldsymbol{e_z}.$$

Using the cross-product relations for the basis vectors we then find:

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{v} &= \Big((\boldsymbol{\nabla}v_\rho)_z\,\boldsymbol{e_\theta} - (\boldsymbol{\nabla}v_\rho)_\theta\,\boldsymbol{e_z}\Big) + \Big(\frac{1}{\rho}(\boldsymbol{\nabla}\rho v_\theta)_\rho\,\boldsymbol{e_z} - \frac{1}{\rho}(\boldsymbol{\nabla}\rho v_\theta)_z\,\boldsymbol{e_\rho}\Big) + \Big((\boldsymbol{\nabla}v_z)_\theta\,\boldsymbol{e_\rho} - (\boldsymbol{\nabla}v_z)_\rho\,\boldsymbol{e_\theta}\Big)\\ &= \Big(\frac{1}{\rho}\frac{\partial v_z}{\partial\theta} - \frac{\partial v_\theta}{\partial z}\Big)\boldsymbol{e_\rho} + \Big(\frac{\partial v_\rho}{\partial z} - \frac{\partial v_z}{\partial\rho}\Big)\boldsymbol{e_\theta} + \Big(\frac{1}{\rho}\frac{\partial(\rho v_\theta)}{\partial\rho} - \frac{1}{\rho}\frac{\partial v_\rho}{\partial\theta}\Big)\boldsymbol{e_z}. \qquad \text{(C.4)}\end{aligned}$$

This is a complicated expression, but one way to keep a rough track of the form of the terms is to notice that each time a θ appears, either in v_θ or in $\frac{\partial}{\partial\theta}$ then the ρ scaling factor also appears in the following fashion $\frac{1}{\rho}\frac{\partial(\rho v_\theta)}{\partial\rho}$ and $\frac{1}{\rho}\frac{\partial(\rho v_\theta)}{\partial z} = \frac{\partial v_\theta}{\partial z}$, or as $\frac{1}{\rho}\frac{\partial}{\partial\theta}$. Apart from the appearance of the inserted scaling factors the expression has the same antisymmetric form of the familiar cross-product in Cartesian coordinates.

C.2 Spherical Coordinates

We define our spherical coordinates (r, θ, ϕ) in relation to Cartesian coordinates[1] by

$$\begin{aligned} x &= r\cos\phi\cos\theta \\ y &= r\cos\phi\sin\theta \\ z &= r\sin\phi \end{aligned} \tag{C.5}$$

where the $\theta \in [0, 2\pi)$ and $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. By following the methods outlined for cylindrical coordinates we may derive that

$$\boldsymbol{e_r} = \begin{pmatrix} \cos\phi\cos\theta \\ \cos\phi\sin\theta \\ \sin\phi \end{pmatrix}, \quad \boldsymbol{e_\theta} = \begin{pmatrix} -\sin\theta \\ \cos\theta \\ 0 \end{pmatrix}, \quad \boldsymbol{e_\phi} = \begin{pmatrix} -\sin\phi\cos\theta \\ -\sin\phi\sin\theta \\ \cos\phi \end{pmatrix}.$$

Hence,

$$\begin{aligned} \boldsymbol{e_x} &= (\cos\theta\cos\phi)\,\boldsymbol{e_r} - (\sin\theta)\,\boldsymbol{e_\theta} - (\cos\theta\sin\phi)\,\boldsymbol{e_\phi} \\ \boldsymbol{e_y} &= (\sin\theta\cos\phi)\,\boldsymbol{e_r} + (\cos\theta)\,\boldsymbol{e_\theta} - (\sin\theta\sin\phi)\,\boldsymbol{e_\phi} \\ \boldsymbol{e_z} &= (\sin\phi)\,\boldsymbol{e_r} + (\cos\phi)\,\boldsymbol{e_\phi} \end{aligned}$$

The infinitesimal displacements, along a path where only one of each coordinate varies, are $dl_r = dr$, $dl_\theta = r\cos\phi d\theta$ and $dl_\phi = r d\phi$ and the volume element is $dv = r^2\cos\phi dr\, d\theta\, d\phi$. For fixed radius r the infinitesimal area is $da = r^2\cos\phi\, d\theta\, d\phi$. Hence we have

$$\boldsymbol{\nabla} = \frac{\partial}{\partial r}\boldsymbol{e_r} + \frac{1}{r\cos\phi}\frac{\partial}{\partial\theta}\boldsymbol{e_\theta} + \frac{1}{r}\frac{\partial}{\partial\phi}\boldsymbol{e_\phi}. \tag{C.6}$$

so that the gradient of a scalar function f is

$$\boldsymbol{\nabla} f = \frac{\partial f}{\partial r}\boldsymbol{e_r} + \frac{1}{r\cos\phi}\frac{\partial f}{\partial\theta}\boldsymbol{e_\theta} + \frac{1}{r}\frac{\partial f}{\partial\phi}\boldsymbol{e_\phi}. \tag{C.7}$$

To compute the divergence and the curl in spherical coordinates we first note that

$$\boldsymbol{e_r}\times\boldsymbol{e_\theta} = \boldsymbol{e_\phi}, \quad \boldsymbol{e_\theta}\times\boldsymbol{e_\phi} = \boldsymbol{e_r}, \quad \text{and} \quad \boldsymbol{e_\phi}\times\boldsymbol{e_r} = \boldsymbol{e_\theta}$$

[1]Note that there are two standard ways to define spherical coordinates, in this definition the angles θ and ϕ are both measured at the origin, with θ being measured in the xy-plane from the x-axis, and ϕ being measured from the xy-plane towards the positive part of the z-axis. Another common choice of the angles is to replace ϕ by $\frac{\pi}{2} - \phi$, or to re-order the labelling of the angle variables. For example to compare with the conventions in Griffiths whose coordinates we label by (r_G, θ_G, ϕ_G), one uses $r_G = r$, $\phi_G = \theta$ and $\theta_G = \frac{\pi}{2} - \phi$.

as direct computation will confirm. Using $\boldsymbol{\nabla} r = \boldsymbol{e_r}$, $\boldsymbol{\nabla}\theta = \frac{1}{r\cos\phi}\,\boldsymbol{e_\theta}$, and $\boldsymbol{\nabla}\phi = \frac{1}{r}\,\boldsymbol{e_\phi}$. By taking cross-products of these vectors, we may construct the divergenceless vectors:

$$\frac{1}{r\cos\phi}\,\boldsymbol{e_\phi}, \quad \frac{1}{r}\,\boldsymbol{e_\theta}, \quad \text{and} \quad \frac{1}{r^2\cos\phi}\,\boldsymbol{e_r}.$$

By expanding a vector $\boldsymbol{v}$ in terms of these divergenceless vectors as

$$\boldsymbol{v} = (v_r r^2\cos\phi)\Big(\frac{1}{r^2\cos\phi}\,\boldsymbol{e_r}\Big) + (v_\theta r)\Big(\frac{1}{r}\,\boldsymbol{e_\theta}\Big) + (v_\phi r^2\cos\phi)\Big(\frac{1}{r^2\cos\phi}\,\boldsymbol{e_\phi}\Big)$$

we simplify the computation of the divergence. Now

$$\begin{aligned}\boldsymbol{\nabla}\cdot\boldsymbol{v} &= (\boldsymbol{\nabla}(v_r r^2\cos\phi))\cdot\Big(\frac{1}{r^2\cos\phi}\,\boldsymbol{e_r}\Big) + (\boldsymbol{\nabla}(v_\theta r))\cdot\Big(\frac{1}{r}\,\boldsymbol{e_\theta}\Big) + (\boldsymbol{\nabla}(v_\phi r^2\cos\phi))\cdot\Big(\frac{1}{r^2\cos\phi}\,\boldsymbol{e_\phi}\Big)\\ &= \frac{1}{r^2}\frac{\partial}{\partial r}(r^2\,v_r) + \frac{1}{r\cos\phi}\frac{\partial}{\partial\theta}(v_\theta) + \frac{1}{r\cos\phi}\frac{\partial}{\partial\phi}(\cos\phi\,v_\phi).\end{aligned}$$

To construct the curl we use the curl-free vectors:

$$\boldsymbol{\nabla} r = \boldsymbol{e_r}, \quad \boldsymbol{\nabla}\theta = \frac{1}{r\cos\phi}\,\boldsymbol{e_\theta}, \quad \text{and} \quad \boldsymbol{\nabla}\phi = \frac{1}{r}\,\boldsymbol{e_\phi}$$

to expand a general vector $\boldsymbol{v}$ as

$$\boldsymbol{v} = v_r\,\boldsymbol{e_r} + v_\theta r\cos\phi\Big(\frac{1}{r\cos\phi}\,\boldsymbol{e_\theta}\Big) + v_\phi r\Big(\frac{1}{r}\,\boldsymbol{e_\phi}\Big)$$

then we find

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{v} &= (\boldsymbol{\nabla}v_r)\times\boldsymbol{e_r} + \frac{1}{r\cos\phi}(\boldsymbol{\nabla}(v_\theta r\cos\phi))\times\boldsymbol{e_\theta} + \frac{1}{r}(\boldsymbol{\nabla}(v_\phi r))\times\boldsymbol{e_\phi}\\ &= \Big((\boldsymbol{\nabla}v_r)_\phi\,\boldsymbol{e_\theta} - (\boldsymbol{\nabla}v_r)_\theta\,\boldsymbol{e_\phi}\Big) + \frac{1}{r\cos\phi}\Big((\boldsymbol{\nabla}(v_\theta r\cos\phi))_r\,\boldsymbol{e_\phi} - (\boldsymbol{\nabla}(v_\theta r\cos\phi))_\phi\,\boldsymbol{e_r}\Big)\\ &\quad + \frac{1}{r}\Big((\boldsymbol{\nabla}(v_\phi r))_\theta\,\boldsymbol{e_r} - (\boldsymbol{\nabla}(v_\phi r))_r\,\boldsymbol{e_\theta}\Big)\\ &= \Big(\frac{1}{r}(\boldsymbol{\nabla}(v_\phi r))_\theta - \frac{1}{r\cos\phi}(\boldsymbol{\nabla}(v_\theta r\cos\phi))_\phi\Big)\boldsymbol{e_r} + \Big((\boldsymbol{\nabla}v_r)_\phi - \frac{1}{r}(\boldsymbol{\nabla}(v_\phi r))_r\Big)\boldsymbol{e_\theta}\\ &\quad + \Big(\frac{1}{r\cos\phi}(\boldsymbol{\nabla}(v_\theta r\cos\phi))_r - (\boldsymbol{\nabla}v_r)_\theta\Big)\boldsymbol{e_\phi}\\ &= \frac{1}{r\cos\phi}\Big(\frac{\partial}{\partial\theta}(v_\phi) - \frac{\partial}{\partial\phi}(v_\theta\cos\phi)\Big)\boldsymbol{e_r} + \frac{1}{r}\Big(\frac{\partial}{\partial\phi}(v_r) - \frac{\partial}{\partial r}(v_\phi r)\Big)\boldsymbol{e_\theta}\\ &\quad + \frac{1}{r}\Big(\frac{\partial}{\partial r}(v_\theta r) - \frac{1}{\cos\phi}\frac{\partial}{\partial\theta}(v_r)\Big)\boldsymbol{e_\phi}.\end{aligned} \tag{C.8}$$

D. Orthogonal Polynomials

The trigonometric functions $\sin(nx)$ and $\cos(ny)$ for $n \in \mathbb{Z}$ may be used as a basis in which other functions can be written. The sense that these functions form a basis comes from the idea that we may define an inner product over pairs of real functions $f(x)$ and $g(x)$ given by an integral:

$$\langle f(x), g(x) \rangle = \int_S f(x)g(x)dx$$

for some suitable interval S. Now when we pick $S = [0, \pi]$ we may show that the trigonometric functions $\sin(nx)$ and $\cos(ny)$ for $n \in \mathbb{Z}$ are orthogonal functions. Let I_{nm} denote the integral

$$I_{nm} = \int_0^\pi \sin(nx)\sin(mx)dx$$

then using integration by parts one can show that

$$I_{nm} = \frac{m}{n}\int_0^\pi \cos(nx)\cos(mx)dx.$$

Using the double-angle formula for cosine we have

$$I_{nm}\Big(\frac{n}{m} - 1\Big) = \int_0^\pi \cos((n+m)x)dx = 0.$$

Hence if $n \neq m$ then $I_{nm} = 0$, while if $n = m$ we have, by direct computation, $I_{nn} = \frac{\pi}{2}$. Therefore $I_{nm} = \delta_{nm}\frac{\pi}{2}$, i.e.

$$\int_0^\pi \sin(nx)\sin(mx)dx = \frac{\pi}{2}\delta_{nm}$$

This is the sense in which $\sin(nx)$ and $\sin(mx)$ for $n \neq m$ are orthogonal: the integral above, interpreted as an inner product is zero. This result is incredibly useful, we may think of the functions $\sin(nx)$ as forming a basis in which all other functions may be expanded[1] (this idea rests upon the notion that the basis is complete, which is the case).

[1]The interested reader is encouraged to look up Fourier series at this point.

We may use the integral to take the equivalent of a dot product against one of these basis functions so as to read off the coefficient of a function expanded in terms of the basis functions. Suppose that we were interested in writing a function $f(x)$ as an expansion of the form $f(x) = \sum_{n=1}^{\infty} \alpha_n \sin(nx)$, then we could compute the coefficients α_m by integrating $f(x)$ against $\sin(mx)$, i.e.

$$\int_0^\pi f(x)\sin(mx)dx = \int_0^\pi \sum_{n=1}^{\infty} \alpha_n \sin(nx)\sin(mx)dx = \sum_{n=1}^{\infty} \alpha_n \delta_{nm}\frac{\pi}{2} = \frac{\pi}{2}\alpha_m.$$

Put another way, we have found that

$$f(x) = \sum_{n=1}^{\infty} \Big(\frac{2}{\pi}\int_0^\pi f(x)\sin(nx)\,dx\Big)\sin(nx).$$

This idea which is useful in electromagnetism and at the heart of Fourier series is part of a wider subject. The rest of this appendix is included for the interested reader.

D.1 Classical Orthogonal Polynomials

It is natural to wonder whether there is a canonical basis of functions which are orthogonal. If so we may expand a function $f(x)$ in such a basis $F_n(x)$, so that

$$f(x) = \sum_n \alpha_n F_n(x) \tag{D.1}$$

and quickly interrogate the coefficients α_n using the orthogonality of the functions F_n:

$$\langle F_m, f(x)\rangle = \sum_n a_n \langle F_m, F_n\rangle = \alpha_m \langle F_m, F_m\rangle \quad \Longrightarrow \quad \alpha_m = \frac{\langle F_m, f(x)\rangle}{\langle F_m, F_m\rangle}. \tag{D.2}$$

But what are the $F_n(x)$ and under which conditions do they exist? We will now follow a very general procedure and develop the classical orthogonal polynomials in a single expression.

Theorem D.1. *Let*

$$F_n(x) = \frac{1}{w(x)}\frac{d^n}{dx^n}(w(x)s^n(x)) \qquad \text{for} \quad n = 0, 1, 2, \ldots \tag{D.3}$$

where $F_1(x)$ is a first order polynomial in x, $s(x)$ is a polynomial in x of degree two or less with only real roots and $w(x)$, the weight function, is a strictly positive function integrable in the region (a, b) satisfying $w(a)s(a) = w(b)s(b) = 0$. Then $F_n(x)$ is a polynomial of degree n in x and

$$\int_a^b p_k(x)F_n(x)w(x)dx = 0 \qquad \forall \quad k < n \tag{D.4}$$

where $p_k(x)$ is any polynomial of degree k.

The formulation of $F_n(x)$ above is sufficiently versatile to include all the classical orthogonal polynomials. By specifying precisely the function $s(x)$ we will be able to determine $w(x)$ and the range (a, b) and for each choice we will find a new series of orthogonal polynomials which will be useful in different settings. These will be described after we have proved the theorem. In order to do this we will make use of two lemmas.

Lemma D.1. *The following is an identity:*

$$\frac{d^m}{dx^m}(ws^n p_{\leq k}) = ws^{n-m} p_{\leq k+m} \qquad \forall \quad m \leq n \tag{D.5}$$

Proof. For $m = 1$ the we have

$$\frac{d}{dx}(ws^n p_{\leq k}) = \frac{dw}{dx} s^n p_{\leq k} + nws^{n-1}\frac{ds}{dx} p_{\leq k} + ws^n p_{\leq k-1} \tag{D.6}$$

where we emphasise that $p_{\leq k}$ is notation for any polynomial of degree $\leq k$, hence $\frac{dp_{\leq k}}{dx} = p_{\leq k-1}$. Looking at the above expression, and considering the lemma, we would like to count the polynomial degree of each of the three terms on the right hand side and check that we can factorise each term into terms of the form $ws^{n-1}p_{\leq k+1}$. We are obstructed in doing this by the $\frac{dw}{dx}$ in the first term. However we can use the definition of the orthogonal polynomials in the main theorem to overcome this obstruction. We have

$$F_1(x) = \frac{1}{w}\frac{d}{dx}(ws) = \frac{1}{w}\frac{dw}{dx}s + \frac{ds}{dx} = p_1(x). \tag{D.7}$$

In the last equality we are simply noting that by definition $F_1(x)$ is a first order polynomial in x. Rearranging we have,

$$s\frac{dw}{dx} = wp_1 - w\frac{ds}{dx} \tag{D.8}$$

which we may use to rewrite equation (D.6) in a way that clearly factorises according to the lemma. Substituting we find

$$\begin{aligned}
\frac{d}{dx}(ws^n p_{\leq k}) &= (wp_1 - w\frac{ds}{dx})s^{n-1}p_{\leq k} + nws^{n-1}\frac{ds}{dx}p_{\leq k} + ws^n p_{\leq k-1} \\
&= wp_1 s^{n-1}p_{\leq k} - w\frac{ds}{dx}s^{n-1}p_{\leq k} + nws^{n-1}\frac{ds}{dx}p_{\leq k} + ws^n p_{\leq k-1} \\
&= ws^{n-1}p_{\leq k+1} + (n-1)ws^{n-1}\frac{ds}{dx}p_{\leq k} + ws^n p_{\leq k-1} \\
&= ws^{n-1}p_{\leq k+1}
\end{aligned} \tag{D.9}$$

where in the last line we have used the fact that $s(x)$ is a polynomial of degree two or less, i.e. $s = p_{\leq 2}$, and hence $\frac{ds}{dx} = p_{\leq 1}$. Taking additional derivatives and repeating the process gives the proof of the lemma. □

We will need a second lemma which is derived from equation D.5 in the lemma above.

Lemma D.2. *All the m derivatives of ws^n vanish at $x = a, b$ for all $m < n$, i.e.*

$$\frac{d^m}{dx^m}(ws^n)\bigg|_{x=a,b} = 0 \qquad \forall \quad m \leq n. \tag{D.10}$$

Proof. From equation D.5, when $k = 0$ we have

$$\frac{d^m}{dx^m}(ws^n) = ws^{n-m}p_{\leq m} \tag{D.11}$$

and at $x = a, b$ we have

$$\frac{d^m}{dx^m}(ws^n)\bigg|_{x=a,b} = (ws)\bigg|_{x=a,b}(s^{n-m-1}p_{\leq m})\bigg|_{x=a,b} = 0. \tag{D.12}$$

□

Armed with this lemma we can now prove the main theorem.

Proof. (Of Theorem D.1)

$$\begin{aligned}
\int_a^b p_k F_n w dx &= \int_a^b p_k \frac{d^n}{dx^n}(ws^n)dx \\
&= \int_a^b p_k \frac{d}{dx}(\frac{d^{n-1}}{dx^{n-1}}(ws^n))dx \\
&= \int_a^b -\frac{dp_k}{dx}\frac{d^{n-1}}{dx^{n-1}}(ws^n)dx + \underbrace{\left[p_k \frac{d^{n-1}}{dx^{n-1}}(ws^n)\right]_a^b}_{=0 \text{ by lemma D.2}}
\end{aligned} \tag{D.13}$$

Repeating the integration by parts k times gives

$$\begin{aligned}
\int_a^b p_k F_n w dx &= \int_a^b (-1)^k \frac{d^k p_k}{dx^k}\frac{d^{n-k}}{dx^{n-k}}(ws^n)dx \\
&= C\int_a^b \frac{d}{dx}(\frac{d^{n-k-1}}{dx^{n-k-1}}(ws^n))dx \\
&= C\left[\frac{d^{n-k-1}}{dx^{n-k-1}}(ws^n)\right]_a^b \\
&= 0
\end{aligned} \tag{D.14}$$

where $C = (-1)^k \frac{d^k p_k}{dx^k}$ is a constant and we have used lemma D.2 again in the final line. To prove that $F_n(x)$ is a polynomial of degree n we use equation D.5 to note that $\frac{d^n}{dx^n}(ws^n) = wp_{\leq n}$ and so

$$F_n(x) = \frac{1}{w}\frac{d^n}{dx^n}(ws^n) = p_{\leq n}. \tag{D.15}$$

To show that F_n is a polynomial of degree n we write

$$F_n(x) = a_n x^n + p_{\leq n-1} \tag{D.16}$$

where a_n is a constant and we note that the repeated index does not indicate the summation convention here. Now consider the positive definite integral (and use the power of the orthogonal polynomials):

$$\begin{aligned} h_n \equiv \int_a^b F_n^2 w dx &= \int_a^b F_n(a_n x^n + p_{\leq n-1}) w\, dx \\ &= \int_a^b F_n a_n x^n w\, dx + \underbrace{\int_a^b F_n p_{\leq n-1} w\, dx}_{=0} \end{aligned} \tag{D.17}$$

where we have used the second part of the theorem (which we have proved first) to see that the second integral vanishes. Consequently as the left hand side is positive definite then $a_n \neq 0$ and so $F_n(x) = p_n$. □

The classical orthogonal polynomials were not found as presented above. They were each found individually as solutions to particular differential equations. Consequently the historical definitions of orthogonal polynomials are all normalised individually with a factor of $\frac{1}{K_n}$, hence we will now adopt the normalised function as our definition of the orthogonal polynomial $F_n(x)$:

$$F_n(x) \equiv \frac{1}{K_n w}\frac{d^n}{dx^n}(ws^n). \tag{D.18}$$

The expression above is called the generalised Rodrigues formula after (Benjamin) Olinde Rodrigues, a Frenchman who first wrote down this formula in 1816. It was later independently discovered by Sir James Ivory in 1824 and Karl Jacobi in 1827. After writing down this formula Rodrigues became a banker. Some time later Rodrigues contributed again to mathematics and has a strong claim to discovering the quaternions three years before Hamilton and without the use of a bridge.

In table D.1 we list the orthogonal polynomials and the classification has been organised according to the degree of $s(x)$. It is important to note the range of the integration where these polynomials are orthogonal, this gives a good indication of where these polynomials may be useful. For example, the Hermite polynomials, are useful in Hilbert space, the Laguerre

Name of $F_n(x)$	$s(x)$	$w(x)$	$[a,b]$
The Hermite polynomials, H_n.	p_0	e^{-x^2}	$[-\infty,+\infty]$
The Laguerre polynomials, $L_n^\nu(x)$.	x	$x^\nu e^{-x}$, $\nu > -1$	$[0,+\infty]$
The Jacobi polynomials, $P_n^{\mu,\nu}(x)$.	$1-x^2$	$(1-x)^\mu(1-x)^\nu$, $\mu,\nu > -1$	$[-1,+1]$

Table D.1: The classical series of orthogonal polynomials.

polynomials are useful when separating and solving the radial part of the Schrödinger equation for hydrogen, and the Jacobi polynomials occur in the study of rotation groups. The Laguerre and Jacobi polynomial series are parameterised by ν and (μ,ν) respectively and specific values of these parameters select out interesting sub-series of the orthogonal polynomials. The most famous case of this are the Legendre polynomials which are a sub-class of the Jacobi polynomials having $\mu = 0$, $\nu = 0$ and hence $w(x) = 1$. The Legendre polynomials occur when finding solutions of the Laplace equation in spherical coordinates (using the method of separation of variables), consequently they are closely related to spherical harmonics. The Legendre polynomials occur naturally in multipole expansions, recall this is the expansion resulting from translation of a spherically symmetric point source and the multipole expansion can be usefully rewritten as a sum of Legendre polynomials.

D.2 The Legendre Polynomials

The Legendre polynomials are the orthogonal polynomials of greatest relevance to electromagnetism. In this section we collect a few facts about these functions.

The Legendre polynomials, up to normalisation, are the Jacobi polynomials defined in table D.1 with $\mu = 0$ and $\nu = 0$, hence $w(x) = 1$. Then the Legendre polynomials are

$$P_l(x) = \frac{1}{2^l l!}\frac{d^l}{dx^l}((x^2-1)^l) = \frac{1}{K_l}F_l \qquad \text{where } l \in \{0,1,2,3,4,\ldots\}$$

i.e. the normalisation constant is $K_l = (-1)^l 2^l l!$. This is chosen so that $P_l(1) = 1$. The first few Legendre polynomials are listed in table D.2 In general we have

$$\int_{-1}^{1} P_n(x)P_m(x)\,dx = \frac{2}{2n+1}\delta_{nm}.$$

By construction we also know that

$$\int_{-1}^{1} p_k(x)P_l(x)\,dx = 0 \qquad \forall \qquad k < l$$

l	$P_l(x)$	$\int_{-1}^{1}(P_l(x))^2\,dx$
0	1	2
1	x	$\frac{2}{3}$
2	$\frac{3x^2-1}{2}$	$\frac{2}{5}$
3	$\frac{5x^3-3x}{2}$	$\frac{2}{7}$
4	$\frac{35x^4-30x^2+3}{8}$	$\frac{2}{9}$
5	$\frac{63x^5-70x^3+15x}{8}$	$\frac{2}{11}$

Table D.2: The Legendre polynomials $P_l(x)$ for $l \leq 5$

where $p_k(x)$ is any polynomial of degree k. The Legendre polynomials form a complete basis of the polynomial functions so that any polynomial of degree n, $p_n(x)$, can be expressed as

$$p_n(x) = \sum_{l=0}^{n} C_l P_l$$

where $l \leq n$ and C_l are constant that may be determined by taking integrals of $p_n(x)$ against the Legendre polynomials and using the orthogonality conditions:

$$\int_{-1}^{1} p_n(x) P_k(x)\,dx = \sum_{l=0}^{n} C_l \int_{-1}^{1} P_k P_l\,dx = \frac{2}{2k+1} C_k$$

The Legendre polynomials satisfy the recursion relation

$$(l+1)P_{l+1}(x) = (2l+1)xP_l(x) - lP_{l-1}(x) \qquad \text{where } l \in \{0,1,2,3,4,\ldots\}$$

which give an alternative way to construct the Legendre polynomials once one defines $P_0(x) = 1$ and $P_1(x) = x$. Another way to recover the Legendre polynomials is as the solutions to Legendre's equation:

$$\frac{d}{dx}\Big((1-x^2)\frac{dP_l}{dx}\Big) = -l(l+1)P_l \qquad \text{where } l \in \{0,1,2,3,4,\ldots\}$$

it is this last property of the Legendre polynomials which make them so useful in solving Laplace's equation in spherical coordinates.

E. Tutorial Exercises

SPECIAL RELATIVITY AND ELECTROMAGNETISM PROBLEMS I

In this tutorial you will practise working with index notation and the Einstein summation convention; and vector calculus. Questions marked with $\star$ are the most crucial problems to study each week.

$\star$ 1. Let f be a scalar field, and let $\boldsymbol{u}$ and $\boldsymbol{v}$ be vector fields. Use index notation to prove the following product rules:

(a) $\boldsymbol{\nabla} \cdot (f\boldsymbol{v}) = f\boldsymbol{\nabla} \cdot \boldsymbol{v} + \boldsymbol{v} \cdot \boldsymbol{\nabla} f$

(b) $\boldsymbol{\nabla} \cdot (\boldsymbol{u} \times \boldsymbol{v}) = \boldsymbol{v} \cdot (\boldsymbol{\nabla} \times \boldsymbol{u}) - \boldsymbol{u} \cdot (\boldsymbol{\nabla} \times \boldsymbol{v})$

(c) $\boldsymbol{\nabla} \times (f\boldsymbol{v}) = \boldsymbol{\nabla} f \times \boldsymbol{v} + f(\boldsymbol{\nabla} \times \boldsymbol{v})$

(d) $\boldsymbol{\nabla} \times (\boldsymbol{u} \times \boldsymbol{v}) = (\boldsymbol{v} \cdot \boldsymbol{\nabla})\boldsymbol{u} - (\boldsymbol{u} \cdot \boldsymbol{\nabla})\boldsymbol{v} + \boldsymbol{u}(\boldsymbol{\nabla} \cdot \boldsymbol{v}) - \boldsymbol{v}(\boldsymbol{\nabla} \cdot \boldsymbol{u})$

$\star$ 2. Prove the following vector identities using index notation:

(a) $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{v}) = 0$,

(b) $\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{v}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{v}) - \boldsymbol{\nabla}^2 \boldsymbol{v}$.

(Hint: For the last part use the identity: $\varepsilon_{ijk}\, \varepsilon^{klm} = \delta_i^l\, \delta_j^m - \delta_i^m\, \delta_j^l$).

3. Calculate the line integral of the vector function $\boldsymbol{v} = \begin{pmatrix} x \\ y \\ xyz \end{pmatrix}$ from the origin $\mathbf{0}$ to the point $\boldsymbol{a} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ when the path P is

(a) The straight line along the x-axis joining $\mathbf{0}$ to $\boldsymbol{a}$.

(b) The path of straight lines: $\mathbf{0} \to \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \to \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \to \boldsymbol{a}$.

(c) The path of straight lines: $\mathbf{0} \to \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \to \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \to \boldsymbol{a}$.

What is the line integral of the closed loop that travels along the path in (a) and returns along the path in (b) in reverse? How does this compare with the line integral around the closed path that starts with the path in part (a) and returns to $\mathbf{0}$ along the path in part (c) in reverse?

4. Compute the surface integral of the vector field $\boldsymbol{v} = \frac{1}{R}\boldsymbol{r}$, where $\boldsymbol{r}$ is the radial vector over

   (a) the circle of radius R in two dimensions; and

   (b) the surface of the sphere of radius R in three dimensions.

5. Compute the volume integral of the scalar field $f = \sqrt{x^2 + y^2}$ over the volume bounded by the sphere of radius R centred at the origin in $\mathbb{R}^3$.

SPECIAL RELATIVITY AND ELECTROMAGNETISM PROBLEMS II

In this tutorial you will continue to revise vector calculus and work with the fundamental theorems of vector calculus, in particular Green's theorem and Stokes' theorem. You will also begin working with electrostatics using Coulomb's law and some of its consequences. Questions marked with $\star$ are the most crucial problems to study each week.

$\star$ 6. In the following let f, g be scalar functions, let $\boldsymbol{u}$ and $\boldsymbol{v}$ be vector functions and let $\boldsymbol{k}$ denote a constant vector. Use the fundamental theorems for the gradient, divergence and curl[1] to derive the following, for an appropriate volume V, surface S and path P,

(a) $\int_V (\boldsymbol{\nabla} f) dv = \oint_S f d\boldsymbol{a}$ [Hint: use $\boldsymbol{u} = \boldsymbol{k} f$ in Green's theorem],

(b) $\int_V (\boldsymbol{\nabla} \times \boldsymbol{v}) dv = -\oint_s \boldsymbol{v} \times d\boldsymbol{a}$ [Hint: Substitute $\boldsymbol{u} = \boldsymbol{v} \times \boldsymbol{k}$ in Green's theorem],

(c) Green's identity: $\int_V \Big(f(\boldsymbol{\nabla}^2 g) + (\boldsymbol{\nabla} f) \cdot (\boldsymbol{\nabla} g) \Big) dv = \oint_S (f \boldsymbol{\nabla} g) \cdot d\boldsymbol{a}$,

(d) Green's second identity: $\int_V (f \boldsymbol{\nabla}^2 g - g \boldsymbol{\nabla}^2 f) dv = \oint_s (f \boldsymbol{\nabla} g - g \boldsymbol{\nabla} f) \cdot d\boldsymbol{a}$, and

(e) $\int_S \boldsymbol{\nabla} f \times d\boldsymbol{a} = -\oint_P f d\boldsymbol{l}$.

$\star$ 7. Compute the divergence of the function $\boldsymbol{v} = \frac{1}{r} \boldsymbol{e_r}$, where r is the length of the three-dimensional radial vector $\boldsymbol{r}$ and $\boldsymbol{e_r} = \frac{\boldsymbol{r}}{r}$. Does your result agree with Green's theorem? Compare your result with the discussion of the divergence of $\frac{1}{r^2} \boldsymbol{e_r}$ in the lecture notes. Is there a Dirac delta function at the origin for this function? Finally, what is the general formula for the divergence of $r^n \boldsymbol{e_r}$ where $n \in \mathbb{Z}$ for $n > -2$ and for which values of n does this divergence require a delta function at the origin?

8. Let $\boldsymbol{F}$ be a vector field in $\mathbb{R}^3$. The following statements are all equivalent:

(a) $\boldsymbol{\nabla} \times \boldsymbol{F} = \boldsymbol{0}$ everywhere,

(b) $\oint \boldsymbol{F} \cdot d\boldsymbol{l} = 0$ for any closed path,

(c) $\int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{F} \cdot d\boldsymbol{l}$ is path-independent, and

(d) $\boldsymbol{F} = -\boldsymbol{\nabla} V$ for some scalar function V.

[1] The latter two of these are known as, Green's theorem and Stokes' theorem respectively.

Prove this claim by showing that (a) is true if and only if (b) is true, (b) is true if and only (c) is true, and (c) is true if and only if (d) is true.

9. Let $\boldsymbol{F}$ be a vector field in $\mathbb{R}^3$. The following statements are all equivalent:

   (a) $\boldsymbol{\nabla} \cdot \boldsymbol{F} = 0$ everywhere,

   (b) $\oint \boldsymbol{F} \cdot d\boldsymbol{a} = 0$ for any closed surface,

   (c) $\int_S \boldsymbol{F} \cdot d\boldsymbol{a}$ is independent of S for any given boundary line, and

   (d) $\boldsymbol{F} = \boldsymbol{\nabla} \times \boldsymbol{A}$ for some vector function $\boldsymbol{A}$.

   Prove this claim by showing that (a) is true if and only if (b) is true, (b) is true if and only (c) is true, and (c) is true if and only if (d) is true.

10. You are given an electric field of the following type $\boldsymbol{E}(\boldsymbol{r}) = q\boldsymbol{r}$

    (a) Find $\boldsymbol{\nabla} \cdot \boldsymbol{E}$,

    (b) Compute $\int\limits_V (\boldsymbol{\nabla} \cdot \boldsymbol{E})\, dv$, where V is a sphere of radius R centred at the origin,

    (c) Compute $\oint_S \boldsymbol{E} \cdot d\boldsymbol{a}$ for a spherical surface of radius R centred at the origin, and

    (d) Compute $\int\limits_\gamma \boldsymbol{E} \cdot \boldsymbol{dl}$ for the curve $\gamma(t) = (t, a, 0)$ and $t \in [0, b]$.

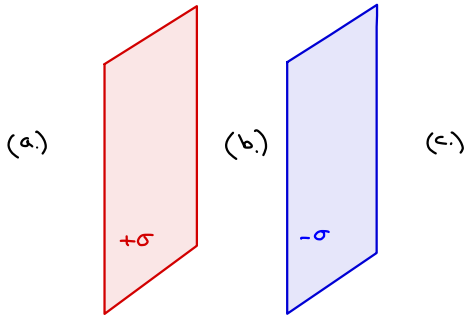
    Do you get the same result in parts (b) and (c)? Why?

SPECIAL RELATIVITY AND ELECTROMAGNETISM PROBLEMS III

In this tutorial you will practise working with Gauss's law to find $\boldsymbol{E}(\boldsymbol{r})$ using an appropriate Gaussian surface, and will start working with the electric potential Φ. Questions marked with $\star$ are the most crucial problems to study each week. **This week you should attempt problem 15 completely and submit your answer to your tutor, it will be marked as if it were an exam problem and returned to you.**

$\star$ 11. A long cylinder carries a charge density[2] that is proportional to the distance from its axis given by $\sigma = k\rho$, where $k \in \mathbb{R}$ and ρ is the radial cylindrical coordinate. Find the electric field *inside* this cylinder.

$\star$ 12. Two infinite parallel planes carry equal but opposite uniform charge densities $\pm\sigma$, see the sketch below, the planes are separated by an unknown distance in the x-direction.



Without calculating any integrals, use the result of example 3.3 to find the electric field in the three regions:

(a) To the left of both planes.

(b) In between both planes.

(c) To the right of both planes.

[2] In cylindrical coordinates we denote the radial coordinate by ρ, so that here, unusually, we will use σ to denote the charge density *per unit volume*.

[Hint: Use the superposition of electric fields.]

13. If the electric field (given in spherical coordinates) is

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{k}{r}\Big(3\,\boldsymbol{e_r} + \cos\phi\cos\theta\,\boldsymbol{e_\theta} + 2\cos\phi\sin\phi\sin\theta\,\boldsymbol{e_\phi}\Big)$$

in some region, for some constant k, what is the charge density?

[Hint: The answer is $\rho(\boldsymbol{r}) = \frac{k\epsilon_0}{r^2}\Big(3 + \sin\theta(1 - 6\sin^2\phi)\Big)$]

⋆ 14. Find the electric potential Φ for the following charge distributions:

(a) A spherical shell of radius R, centred at the origin, and carrying a uniform distribution of charge of constant charge density σ.

(b) An infinitely long, thin wire aligned along the z-axis and carrying a linear charge density of λ. For this case you will not be able to set the zero of the potential to be at infinity (why not?), instead pick an alternative reference point for the potential that gives a finite answer.

(c) A charged infinite plane, coincident with the xy-plane, carrying uniform charge density σ. As in part(b) you will need to pick a reference point for your potential that is not at infinity.

15. Consider the Yukawa potential[3] which is given by

$$\Phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\frac{qe^{-\mu r}}{r}$$

where $\mu \in \mathbb{R}$ and $\mu \geq 0$. Note that for $\mu = 0$ the reduces to the potential for a single source charge q.

(a) Find the electric field $\boldsymbol{E}(\boldsymbol{r})$ associated to this potential.

(b) Use Gauss's law to calculate the charge density that produces the Yukawa potential.

[3]Named after Hideki Yukawa (1907-1981) who proposed this potential in a 1935 paper to resolve problems with an atomic model.

SPECIAL RELATIVITY AND ELECTROMAGNETISM PROBLEMS IV

In this tutorial you will work with a number of standard electrostatic problems with a focus on using the potential, and, where possible, the method of images. Questions marked with $\star$ are the most crucial problems to study each week.

16. Consider a uniformly charged thin sphere with radius R and charge density σ. The sphere has a small hole at the north pole with radius $\frac{d}{2}$ ($d \ll R$). Find the electric field everywhere inside the sphere (but not too close to the hole). [Hint: You only need to solve this problem approximately, so you should not do any integrals here!]

$\star$ 17. Compute the average potential over a sphere of radius R due to a single point charge q located outside the sphere. Confirm that your result is equal to the potential due to q if it were located at the centre of the sphere. [Hint: For simplicity centre the sphere at the origin and place the source charge q on the z-axis.]

18. Give a short argument to justify Earnshaw's Theorem: 'A collection of charged particles cannot be held in a stable equilibrium by electrostatic forces alone.' [Hint: Test the requirements of stable equilibrium using Gauss's law.]

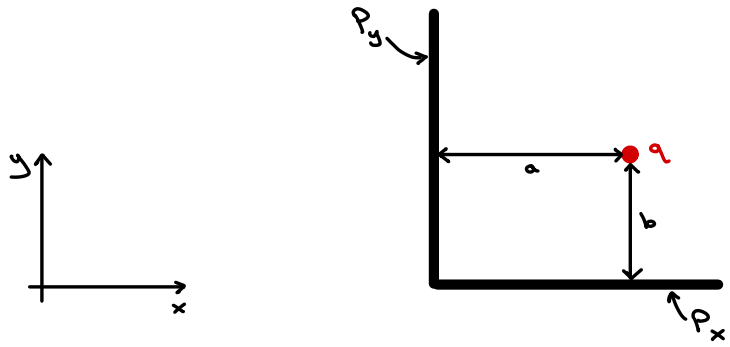
$\star$ 19. A point charge q is located at $\begin{pmatrix} 0 \\ 0 \\ a \end{pmatrix}$, outside a grounded, conducting sphere of radius R centred at the origin.

(a) Explain why the potential outside the sphere may be found by considering the problem of the potential for two charges: q located at $\begin{pmatrix} 0 \\ 0 \\ a \end{pmatrix}$ and a second charge $-\frac{R}{a}q$ located at $\begin{pmatrix} 0 \\ 0 \\ \frac{R^2}{a} \end{pmatrix}$.

(b) Find the potential outside the grounded, conducting sphere.

(c) Find the total induced charge on the surface of the grounded, conducting sphere.

20. Use the method of images to compute the potential due to a point charge q at $\begin{pmatrix} a \\ b \\ 0 \end{pmatrix}$ surrounded by two grounded, conducting (semi-)infinite planes $P_x = \{ \begin{pmatrix} x \\ 0 \\ z \end{pmatrix} \mid x > 0, z \in \mathbb{R} \}$ and $P_y = \{ \begin{pmatrix} 0 \\ y \\ z \end{pmatrix} \mid y > 0, z \in \mathbb{R} \}$. The situation is depicted in cross-section in the plane $z = 0$ from above (looking down the z-axis) by

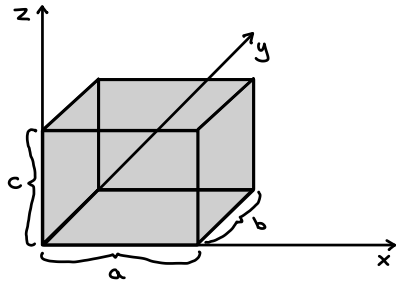


[Hint: You should add more than one image charge.]

SPECIAL RELATIVITY AND ELECTROMAGNETISM PROBLEMS V

In this problem set you will solve Laplace's and Poisson's equations using separation of variables and Green's functions. Questions marked with $\star$ are the most crucial problems to study each week.

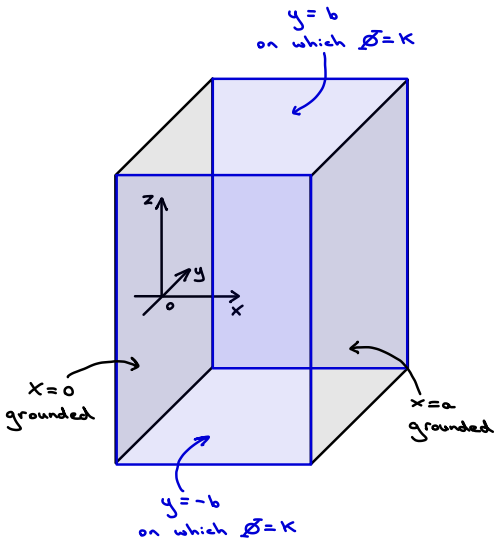
$\star$ 21. Consider a conducting rectangular box of sides a, b and c oriented as shown below



All faces of the box are grounded except the top face (which lies in the plane $z = c$). On the top face the potential is $f(x, y)$. Determine the potential inside the rectangular box in terms of $f(x, y)$.

22. Consider a rectangular pipe constructed along the z-axis and formed out of two infinitely long grounded conductors at $x = 0$ and $x = a$ and another two infinitely long conductors

at $y = -b$ and $y = b$ which are held at constant potential K.



Find the potential within the rectangular pipe.

⋆ 23. Consider an infinitesimally thin disc defined by the points in $\mathbb{R}^3$ satisfying $x^2 + y^2 < R^2,\ z = 0$ carrying a constant surface charge density denoted σ.

(a) Write down the charge density $\rho(\boldsymbol{r})$ for $\boldsymbol{r} \in \mathbb{R}^3$.

[Hint: You will need to use a step function.]

(b) Find the potential along the z axis.

(c) What happens to the potential along the z-axis when $R \longrightarrow \infty$? Compare the result with the formula for the infinite plane.

24. Consider a hemisphere given by $|\boldsymbol{r}| = R$ and $-R < z < 0$ and assume that it has a uniform charge density σ on its surface.

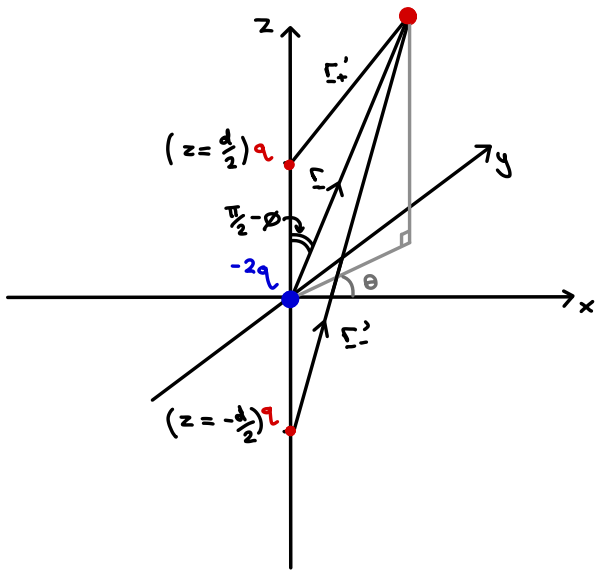
(a) Write the charge density $\sigma(\boldsymbol{r})$ in spherical coordinates.

(b) Evaluate the scalar potential along the z axis.

(c) Find the electric field at the origin.

25. Prove that the Green's function for the two-dimensional Laplace operator is $\frac{1}{2\pi}\ln(|\boldsymbol{r} - \boldsymbol{r'}|)$.

26. Consider the linear electric quadrupole:



consisting of three point charges on the z-axis q at $z = \frac{d}{2}$, $-2q$ at $z = 0$, and q at $z = -\frac{d}{2}$. Find its potential $\Phi(x, y, z)$ and show that for $r \gg d$ it is approximated by the quadrupole term in the multipole expansion for $\Phi(\boldsymbol{r})$.

SPECIAL RELATIVITY AND ELECTROMAGNETISM PROBLEMS VI

In this problem set you will continue to work on electrostatics, but also commence work with magnetostatics including the Lorentz force law, the Biot-Savart law and the magnetic vector potential. Questions marked with ⋆ are the most crucial problems to study each week.

⋆ 27. Find the electric dipole moment of a thin sphere with radius R. The surface charge density σ depends on the spherical coordinate ϕ the following way: $\sigma(\phi) = \sigma_0 \sin\phi$, where σ_0 is a real constant. What is the monopole term in the multipole expansion of this system?

28. Consider two conducting spheres of radius R_1 and R_2 which are very far from each other and are connected by a thin wire and carry a total charge Q. Ignore the interaction between the two balls. Being conductors, the value of the scalar potential at their surface is equal.

    (a) How is the charge distributed between the two spheres?

    (b) What is the electric field at the surface of each?

    (c) How does this problem explain spark-plugs and lightning rods?

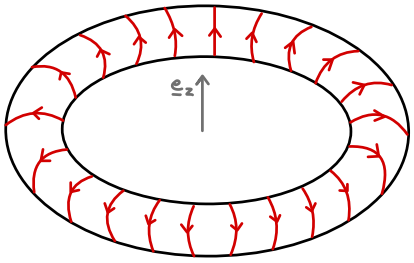
29. Find the trajectory of a test charge Q of mass m moving in a magnetic field $\boldsymbol{B} = B\,\boldsymbol{e_z}$, where B is a constant, whose initial position is $\boldsymbol{r}(0) = \boldsymbol{e_x}$, and whose initial velocity is $\boldsymbol{v}(0) = -\boldsymbol{e_y} + \boldsymbol{e_z}$.

⋆ 30. Find the trajectory of a test charge Q, initially at rest at the origin, subjected to a constant electric field $\boldsymbol{E} = E\,\boldsymbol{e_z}$ and a constant magnetic field $\boldsymbol{B} = B\,\boldsymbol{e_x}$.

31. Consider a thin infinite wire with constant current I, whose position is described by the equations: $x = 0$, $y = 0$. Explain why the current density is $\boldsymbol{J}(x, y, z) = I\delta(x)\delta(y)\,\boldsymbol{e_z}$. Use the current density to find the vector potential (use the integral formula from the lectures). Find the magnetic field from the vector potential.

⋆ 32. Consider a torus around which a long wire is uniformly wrapped (so that you may ap-

proximate each loop of wire by a circle:



where each loop of wire carries a current I and there are n loops of wire in total on the torus. With the torus aligned as shown above (i.e. with the plane of its larger circle being perpendicular to $\boldsymbol{e_z}$). Prove that the magnetic field is in the $\boldsymbol{e_\theta}$ direction (in cylindrical coordinates). Find the magnetic field.

33. A spherical shell of radius R, carrying a uniform surface charge σ is set spinning at constant angular velocity ω. Find the magnetic vector potential produced at a point $\boldsymbol{r}$.

SPECIAL RELATIVITY AND ELECTROMAGNETISM PROBLEMS VII

In this problem set you will continue to work with magnetostatics, although our attention will start to focus on electrodynamics as well, and we will do some work with Faraday's law. Questions marked with $\star$ are the most crucial problems to study each week.

$\star$ 34. Prove that the component of $\boldsymbol{B}$ perpendicular to any surface with a steady surface current is continuous at the surface. Prove that the component of $\boldsymbol{B}$ parallel to any surface with a steady surface current is discontinuous across the surface - and compute the discontinuity.

35. In this problem we will prove a vector identity that is helpful in re-writing the magnetic dipole moment:

$$\oint_P \boldsymbol{e_r} \cdot \boldsymbol{r'}\, d\boldsymbol{l'} = -\,\boldsymbol{e_r} \times \int_S d\boldsymbol{a'}.$$

(a) Use Stokes theorem for $\boldsymbol{v} = \boldsymbol{c}f$ where $\boldsymbol{c}$ is a constant vector and f a scalar function to prove that

$$\int_S \boldsymbol{\nabla} f \times d\boldsymbol{a} = -\oint_P f d\boldsymbol{l}$$

where P is the boundary of S.

(b) Prove that

$$\oint_P (\boldsymbol{c} \cdot \boldsymbol{r}) d\boldsymbol{l} = -\boldsymbol{c} \times \int_S d\boldsymbol{a}$$

where $\boldsymbol{c}$ is a constant vector, by substituting $f = \boldsymbol{c} \cdot \boldsymbol{r}$ into the result of part (a).

(c) Hence prove the identity stated at the start of the question.

36. Find the magnetic dipole moment $\boldsymbol{m}$ of a current I circulating around a square of side L.

$\star$ 37. A square loop of wire with sides of length L whose edges are parallel to the $\boldsymbol{e_x}$ and $\boldsymbol{e_z}$ directions lies with its closest edge at a distance d from an infinitely long straight wire on the z-axis which is carrying a steady current $\boldsymbol{I} = I\,\boldsymbol{e_z}$.

(a) Compute the flux of $\boldsymbol{B}$ through the square of wire.

(b) If the square loop is moved at a velocity $\boldsymbol{v} = v\,\boldsymbol{e_x}$ away from the current what is the electromotive force generated, and in what direction around the square does the current flow?

(c) Repeat part (b) but with $\boldsymbol{v} = v\,\boldsymbol{e_z}$.

⋆ 38. *An alternating current generator:* A square loop of wire with sides of length L lying initially in the plane $y = 0$ and bisected by the z-axis, is rotated at angular speed ω about the z-axis. This wire rotates in a uniform background magnetic field $\boldsymbol{B} = B\,\boldsymbol{e_x}$. Compute the electromotive force $\mathcal{E}(t)$ for this generator.

39. *A transformer:* Two conducting wires are wrapped around a cylinder in such a way that the same magnetic flux passes through every turn of both coils of wire. The first wire has n_1 turns around the cylinder and the second wire has n_2 turns of wire around the cylinder. If there is a changing current $I(t)$ in the first wire show that

$$\frac{\mathcal{E}_2}{\mathcal{E}_1} = \frac{n_2}{n_1}$$

where $\mathcal{E}_2$ is the electromotive force induced in wire two and $\mathcal{E}_1$ is the electromotive force induced in wire one both due to the changing current in wire one.

SPECIAL RELATIVITY AND ELECTROMAGNETISM PROBLEMS VIII

In this problem set we will continue to work with electrodynamics: electromagnetic waves, energy, gauge choices and the retarded and advanced potentials. Questions marked with $\star$ are the most crucial problems to study each week.

40. Prove that the spherical wave
$$f(\boldsymbol{r}, t) = \frac{1}{r} e^{ik(r-ct)}$$
is a solution of the three dimensional wave equation $\Box f = 0$.

41. Prove that there is no non-trivial solution to Maxwell's equation in a vacuum where the electric field is
$$\tilde{\boldsymbol{E}}(\boldsymbol{r}, t) = \tilde{\boldsymbol{E}}_{\boldsymbol{0}} \frac{1}{r} e^{ik(r-ct)}$$
where $\tilde{\boldsymbol{E}}_0$ is a constant.

$\star$ 42. Write down the real electric and magnetic fields for a monochromatic plane wave of amplitude E_0, angular frequency ω that is

(a) Travelling in the direction $-\boldsymbol{e_z}$ and polarised in the $\boldsymbol{e_x}$ direction.

(b) Travelling from the origin to the point $x = 1$, $y = 1$ and $z = 1$ with polarisation parallel to the plane $y = 0$.

$\star$ 43. *The propagation of energy by light:* A monochromatic plane wave in a vacuum has electric field
$$\boldsymbol{E}(x, t) = E_0 \big(\cos(kx - \omega t) \big) \boldsymbol{e_z}.$$

(a) What is the accompanying magnetic field $\boldsymbol{B}(x, t)$?

(b) Compute the energy density $U = \frac{1}{2}\big(\epsilon_0 \boldsymbol{E}^2 + \frac{1}{\mu_0}\boldsymbol{B}^2\big)$ for this wave.

(c) Find the Poynting vector $\boldsymbol{S} = \frac{1}{\mu_0}(\boldsymbol{E} \times \boldsymbol{B})$ for this wave and compare it with U.

44. *(Discussion Question)* Qualitatively discuss the effect on a stationary, isolated (point-like) test charge Q in an electromagnetic vacuum which is impacted by a monochromatic plane wave? (For the purpose of this discussion, you should assume that the electric field of the test charge does not change the electromagnetic wave.)

⋆ 45. Consider the spherical electromagnetic wave with

$$\boldsymbol{E}(\boldsymbol{r},t) = A\frac{\cos\phi}{r}\Big(\cos(u) - \frac{1}{kr}\sin(u)\Big)\,\boldsymbol{e_\theta}$$

where $A \in \mathbb{R}$ is constant and $u = kr - \omega t$ where $c = \frac{\omega}{k}$. Show that $\boldsymbol{E}$ satisfies all four of Maxwell's equations in a vacuum and find the associated magnetic field $\boldsymbol{B}(\boldsymbol{r},t)$.

46. One can solve the inhomogeneous wave equations:

$$\Box\Phi = -\frac{\rho}{\epsilon_0}$$
$$\Box\boldsymbol{\Phi} = -\mu_0\boldsymbol{J}$$

with the retarded and the advanced potentials. Using this ability to solve such equations show that it is always possible to choose the potentials in the Lorenz gauge:

$$\boldsymbol{\nabla}\cdot\boldsymbol{A} = -\epsilon_0\mu_0\frac{\partial\Phi}{\partial t}.$$

SPECIAL RELATIVITY AND ELECTROMAGNETISM PROBLEMS IX

In this problem set we will start to work with special relativity: we will look at simultaneity, time dilations and length contraction; the Lorentz transformation often through some apparent paradoxes. You should aim to submit at least one starred ($\star$) question for marking.

$\star$ 47. *The twin paradox.* Consider two identical twins, called A and B, who have the same age (up to some decimal points at least). Twin B stays on Earth while twin A travels on a spaceship at a very high speed v (relative to B on Earth) to a distant star and back. When the twins are reunited on Earth, are they the same age? If not, which one is younger?

48. Many formulae in special relativity contain the factor $\gamma(v) = \left(\sqrt{1-\frac{v^2}{c^2}}\right)^{-1}$, which controls how strong the deviation from "classical behaviour" is. Here c is the speed of light ($\approx 3 \times 10^8$m/s) and v is the relative velocity of two inertial systems (or the velocity of a particle within a system).

    To get an idea of the size of relativistic effects, compute $\gamma(v)$ for $v = 360$km/h (a very fast train), $v = 30$km/s (earth moving in its orbit around the sun), $v = 0.9c$ and $v = 0.998c$ (velocities common in elementary particle physics).

    Find the speeds v for which $\gamma(v) = 1.01$ (a deviation from classical behaviour by 1%) and $\gamma(v) = 2$ (a deviation by 100%).

$\star$ 49. The muon is an unstable particle that spontaneously decays into an electron and two neutrinos. The mean lifetime of a muon in its rest frame is $\tau = 2.20\mu$s. Most muons that reach the Earth are produced by cosmic rays in the upper atmosphere at altitudes of 15 km with typical speeds of $0.95c$.

    (a) Using Newtonian (non-relativistic) dynamics, what fraction of the muons would reach the Earth without decaying? The decay law is $N = N_0\, e^{-t/\tau}$, where τ is the mean lifetime,

    (b) According to special relativity, what is the mean lifetime of a muon in Earth's frame?

    (c) How far is the Earth from the upper atmosphere for such a (relativistic) muon?

(d) According to special relativity, What fraction of the muons reach the Earth from the upper atmosphere? Think about the problem from both the frame of the Earth and the frame of the muon - do the answers from both frame's agree?

50. Consider two inertial frames: S with coordinates x^μ and S' with coordinates y^μ. The frames are arranged so that the coordinate axes coincide at some initial moment, and in particular x^1 is parallel to y^1.

A rod is stationary in the frame S' where it is measured to have length L' and to make an angle θ' with the y^1 axis. In the frame S the rod (and the frame associated to it, S') is moving with speed v along the x^1 axis. What is the length of the rod for an observer that is stationary in S? What angle does it make with the x^1 axis?

$\star$ 51. Consider a train of length 100m at rest and a tunnel of length 80m. The tunnel has doors on both ends: initially the entry door to the tunnel is open and the exit door is closed, but the doors are programmed so that whenever a train reaches the exit door it is automatically (and immediately) opened to let the train through and the entry door closes automatically as the back of the train passes it.

What speed should the train have so that it would fit inside the tunnel exactly? What about the frame of the train, would people on the train say that it fits inside the tunnel completely? Describe the sequence of events in the train frame and show that there are no contradictions.

SPECIAL RELATIVITY AND ELECTROMAGNETISM PROBLEMS X

In this problem set we will continue our relativistic work, and look at Lorentz transformations, Lorentz invariants and relativistic dynamics. Questions marked with $\star$ are the most crucial problems to study each week.

$\star$ 52. Show that the d'Alembertian operator transforms as a Lorentz scalar under a change in the coordinates given by a Lorentz boost in the $\boldsymbol{e_x}$ direction.

53. Show that if the four-displacement between two events is:

    (a) time-like then there exists an inertial reference frame in which the events occur at the same location,

    (b) space-like then there exists an inertial reference frame in which the events occur at the same time, and

    (c) light-like then the two events could be connected by a light signal.

$\star$ 54. A pion of mass m_π at rest decays into a muon of mass m_μ and a massless neutrino (neutrinos are not really massless, but have negligible mass, and for massless particles, the 4-momentum satisfies $P^\mu P_\mu = 0$).

    (a) Write down the 4-momenta for all the particles.

    (b) Using 4-momentum conservation find the speed of the produced muon.

    (c) On average, how far will the muon travel before disintegrating? (use $m_\pi \simeq 139 MeV/c^2$, $m_\mu \simeq 105 MeV/c^2$, and data about the muon lifetime from the previous problem set.)

$\star$ 55. Using the conservation of 4-momentum show that a single massive particle cannot decay into a single photon[4] (a massless particle which moves at the speed of light), instead it can only decay into two or more photons.

56. There are two basic strategies which are used in particle accelerators. The first has a fixed target, where you smash a particle of mass m_b with energy E_b into a test particle of

[4] A photon is the quantum of the electromagnetic field, and emerges when one quantises electromagnetism in quantum electrodynamics.

mass m_t that is at rest. The second involves colliding beams, where you accelerate two particles to energies E_b for a head-on collision. Either way the goal is to produce some new particle with mass m_X.

(a) Find the mass of the produced particle m_X in terms of the initial data E_b, m_b and m_t for the fixed target scenario.

(b) Find the mass m_X in terms of the beam energies E_b for the colliding beam scenario (you want to produce a particle at rest to make it as massive as possible).

(c) What kind of accelerator would you build if you wanted to discover heavy particles (e.g. the Higgs boson)?