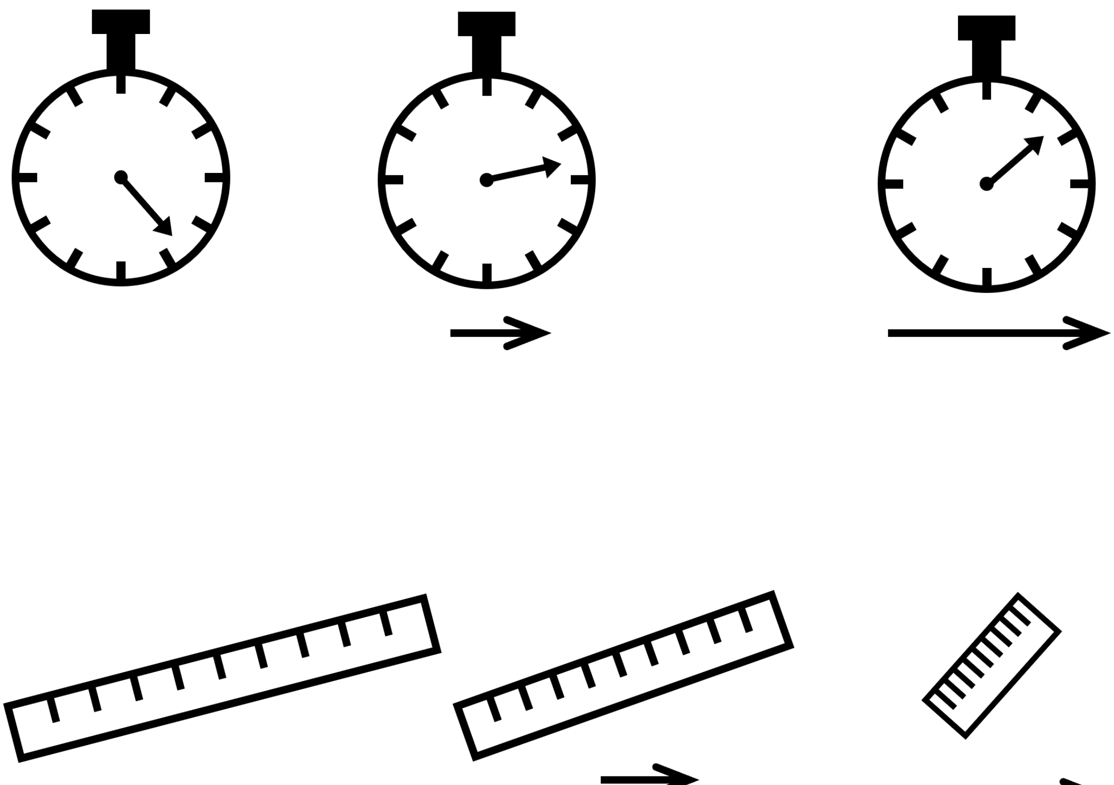
The Special Theory of Relativity



Michael Brownlie

Contents

1 Introduction **1**

1.1 What is relativity and why do we need it? 2

1.2 Key definitions 2

1.3 Einstein's postulates 3

2 Space and Time **4**

2.1 The Lorentz transformations 5

2.2 Can you go faster than light? 9

2.3 Time dilation 10

2.4 Length contraction 13

2.5 Velocity addition and rapidity 16

2.6 Waves and the Doppler effect 17

3 Paradoxes in Special Relativity **24**

3.1 The twins paradox 24

3.2 The barn-pole paradox 29

3.3 Bell's spaceship paradox 32

3.4 Ehrenfest's paradox 33

4 Relativistic Dynamics **34**

4.1 Energy and momentum 35

4.2 Massless particles 38

4.3 Why does $E = \mathcal{M}c^2$? 39

4.4 Pion decay 41

4.5 Proton-electron capture 43

4.6 Compton scattering 44

4.7 Relativistic rockets 46

5 Experimental Evidence for Special Relativity **48**

5.1 Aberration of starlight 48

5.2 The Fizeau experiment 49

5.3 The Michelson–Morley experiment 52

5.4 Atmospheric muons 54

5.5 The Hafele–Keating experiment 55

5.6 Bucherer's experiment 56

5.7 The Cockcroft–Walton experiment 57

6 Relativity as Spacetime Geometry **58**

6.1 A review of Euclidean geometry 59

6.2 Minkowski geometry 63

6.3 Lorentz transformations and 4-vectors 66

6.4 4-velocity and 4-momentum 68

6.5 The wave 4-vector 71

Appendix **73**

1 Introduction

At the end of the nineteenth century it was the widely believed among physicists that light[1] propagates through a medium referred to as the luminiferous aether. However, it proved remarkably difficult to pin down exactly how this aether interacted with matter. Broadly speaking, there were two schools of thought. Augustin-Jean Fresnel proposed that the aether was a more or less stationary fluid which is only weakly affected by material objects moving though it. In contrast, George Stokes suggested that the aether is completely dragged along by matter, in much the same way that the atmosphere, and hence sound waves, is dragged along with the surface of the Earth as it moves. Ordinarily, settling such a dispute would be a simple matter of designing and carrying out a suitably precise measurement. Unfortunately, in this case the experimental results were a rather mixed bag: some supported Stokes, while others aligned more closely with Fresnel's model.

Hendrik Lorentz, with assistance from Henri Poincaré, was able to shed some light on the situation by developing a theory where the aether was completely stationary. He proposed that the reason the aether sometimes appeared as though it was being dragged along by matter was because rulers and clocks interacted with the aether as they moved through it, causing them to measure distances and times incorrectly. Lorentz was able to show that this would indeed be predicted by a stationary aether if the clocks and rulers were governed by electromagnetic interactions. In order to explain the fact that the luminiferous aether was apparently influencing all clocks and rulers, he speculated that all of physics might ultimately be electromagnetic in origin. This approach had a fair amount of success; however, the electromagnetic calculations it required were rather complicated and lengthy.

In 1905, Albert Einstein revolutionised the way physicists looked at this problem in his paper 'On the Electrodynamics of Moving Bodies'. This paper was based on the idea, first put forward by Galileo nearly three hundred years earlier, that it is fundamentally impossible for any inertial observer to measure their velocity through space, except in reference to other material objects. Using only this assumption and mathematics understandable to schoolchildren, Einstein was able to derive in a few pages what had taken his contempories years of blood, sweat, and tears. The value of this approach was further solidified in a second paper, also published that year, 'Does the Inertia of a Body Depend Upon Its Energy Content?', where Einstein used this idea to obtain his most famous result $E = mc^2$. With further developments from others, such as Max Planck and Hermann Minkowski, this new approach to physics became what we now refer to as the special theory of relativity.

These notes present an introduction to the subject of special relativity which more or less follows the approach taken by Einstein in his two 1905 papers, with some slight modifications and additions to reflect a more modern understanding of the subject matter. The only major omission is that we will make no reference to the theory of classical electromagnetism. This reflects the fact that, although it was problems in electromagnetism that motivated the development of the theory, it is actually much more general and wider reaching.

[1] This belief applied equally to all other electromagnetic phenomena and not just light.

1.1 What is relativity and why do we need it?

Technically speaking, special relativity is not really a physical theory in its own right, since it does not contain any description of material objects, which are rather important for modelling the real world. It is perhaps more correct to say that it is a set of conditions which we believe should be obeyed by any reasonable theory, and the study of special relativity is really a study of the things that all such theories must have in common. This makes special relativity applicable in a very general way, and is probably why it has survived to this day, despite the fact that nearly everything we know about the fundamental building blocks of the universe has changed since it was first formulated.

The central principle behind relativity is that, no matter what you do, you simply cannot measure your own velocity. The feeling of motion that we experience in our daily lives is actually a consequence of our relative motion to the ground beneath our feet and the atmosphere around us. This fact will be familiar to anyone who has flown in an aeroplane: despite the nearly $1000\,\mathrm{km\,h^{-1}}$ speed you can sit in your seat as easily as in a lecture theatre, because the air inside the plane is moving just as fast as you.[2] This certainly seems like a reasonable idea, but as with anything in physics there is no real justification for our belief other than the fact that no violation of this principle has been observed in any experiment to date. Nonetheless, special relativity is one of the most well tested theories in all of physics, and so we can feel about as confident in it as it is possible to be.

Even if we accept that the principle of relativity is true, why is it worth investigating? After all, it will already be embedded into all the laws nature anyway, so we'll essentially be passively studying relativity whenever we study anything else. The main reason is that relativity can greatly simplify calculations. For example, it would be essentially impossible to directly calculate how fast a moving pocket watch ticks using fundamental physics; however, with special relativity this can be done in a matter of moments. Furthermore, we can apply relativity to make predictions even in situations where we do not fully understand the physics at play, which is particularly useful if you're trying to investigate new physics.

1.2 Key definitions

Before get started, let us first introduce a few useful definitions.

An event is something which can be described by where and when it occurs. So for example, the beginning and end of a lecture are both events: they occur in a specifiable location and at a specifiable time, both of which can be recorded in a timetable. Within physics we typically encounter events such as the emission of a photon by an excited atom, or the collision between two protons, etc. We call the set of all events at which a particular object is present that object's world line.

[2] This argument also highlights the things about your own motion that you can measure: accelerations. While you don't notice the enormous velocity of the plane, you definitely notice when it accelerated slightly by turbulence.

A reference frame can be thought of as an infinite cubic lattice, with a clock at every vertex and a unit ruler along every edge. For any given observer, there is a special frame, known as the observer's rest frame, in which all of the clocks are at rest relative to that observer. A different frame of reference, that is to say a different lattice of clocks may be translated, rotated, and/or in motion compared to the first, and will assign different descriptions to the same event. We shall denote reference frames with calligraphic font e.g $\mathcal{K}$ and $\mathcal{K}'$.

Coordinates are a set of four numbers which uniquely describe an event. We can use a reference frame $\mathcal{K}$ to produce a natural set of coordinates t, x, y, z, as follows. The value of t for any event is given by the reading on the clock nearest to where the event occurred, while the values of x, y, z refer to the number of rulers which must be traversed in each direction to get from some given origin to the location of the event. A different frame $\mathcal{K}'$ will assign different coordinates t', x', y', z' to the same event.

An inertial frame is a reference frame in which Newton's first law holds true. That is to say, it is a frame relative to which isolated (i.e experiencing no external force) objects move in straight lines at a constant velocity. If an observer's rest frame is inertial, then we know that the observer moves with a constant velocity (of zero) in an inertial frame, and so it must not be moving under the influence of a resultant external force.

1.3 Einstein's postulates

Einstein identified two key postulates from which he could derive all of special relativity. These postulates are:

The principle of relativity, which states that the laws of physics must be identical in any inertial frame of reference. What this means is that, if $\mathcal{K}$ and $\mathcal{K}'$ are two inertial frames, then the algebraic forms of the laws of nature must be unchanged if we replace the coordinates t, x, y, z with their primed counterparts t', x', y', z'.

The light postulate, which states that light (in a vacuum) travels in a straight line at speed c in any inertial frame of reference. In essence, the light postulate is a hypothesis that the propagation of light is determined directly from the fundamental laws of physics, not interaction with a luminiferous aether, and so falls under the purview of the first postulate.

To slow moving creatures, such as us, who are used to the idea that velocities add together linearly these two postulates seem rather at odds with one another. However, as counter intuitive as it seems, the two statements are perfectly compatible. In fact, it turns out that any theory which obeys the first postulate, must possess an invariant speed.[3] Furthermore, all such theories will be virtually identical to special relativity, with the only difference being the replacement of c with the appropriate invariant speed.

[3] The Galilean relativity which we are familiar with from Newtonian mechanics corresponds to the limit as this invariant speed tends towards infinity.

2 Space and Time

One of the most profound implications of special relativity is how it altered the way we think about space and time. As a simple example of why this must happen, let us consider the 'light clock' shown in Fig. 1. In the clock's rest frame $\mathcal{K}'$, the light postulate tells us that the rays bouncing between the two mirrors must move at speed c, and so the time between each tick is

$$\Delta t' = \frac{2h}{c}\,. \tag{2.1}$$

We now consider the same clock, but viewed from a frame $\mathcal{K}$, where it has a velocity of v parallel to the mirrors. Assuming that that the height of the clock doesn't change, the light rays now have slightly further to go in each bounce between the mirrors, but by the light postulate they must still travel at the same speed. Thus, in this frame the time between each tick must by

$$\Delta t = \frac{2\sqrt{h^2 + v^2(\Delta t)^2/4}}{c} \implies \Delta t = \frac{\Delta t'}{\sqrt{1 - v^2/c^2}}\,. \tag{2.2}$$

Note that, although we used the physical system of the light clock to get here, this is simply a statement about the relationship between the coordinates of the two frames for events along the light clock's world line. Thus, the same effect of ticking more slowly in $\mathcal{K}$ when compared to $\mathcal{K}'$ must apply to any clock moving along that same world line.

The only way for us to avoid this conclusion that moving clocks tick more slowly than their stationary counterparts would be if the height of the clock changed between reference frames. Either way, we are forced into the unavoidable conclusion that, if special relativity is true, either distances, times, or both must change between different frames of reference.

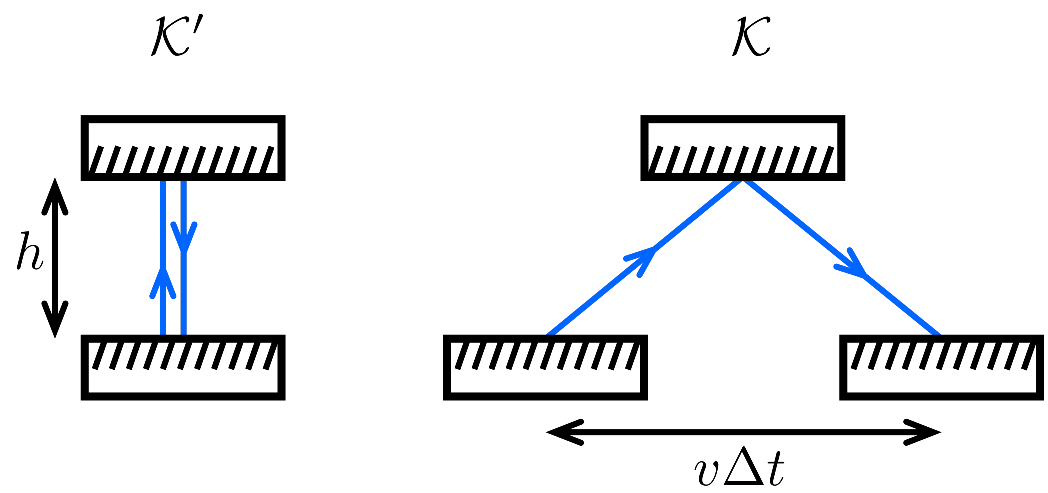


Figure 1. A light clock consisting of two mirrors separated by a height h and a light pulse bouncing between them. Each tick of the clock consists of one round trip of the light pulse between both mirrors.

2.1 The Lorentz transformations

Given two inertial frames $\mathcal{K}$ and $\mathcal{K}'$, we can ask ourselves how the coordinates associated with each frame are related to one another, i.e can we obtain functional relationships which tell us t', x', y', z' for a given event, provided that we know its values of t, x, y, z?

Finding the general form of the coordinate transformation between reference frames is an awful lot of effort for not a lot of reward. For the sake of completeness, a complete description of how to do this is provided in the appendix. Fortunately, a great many problems do not require this full apparatus and can be reduced to the conversion between two frames in the so called 'standard configuration'. Two frames $\mathcal{K}$ and $\mathcal{K}'$ are said to be in the standard configuration if their origins coincide at time $t = t' = 0$, their axes are all aligned with each other, and $\mathcal{K}'$ moves with a velocity v along the x axis in $\mathcal{K}$.

We shall start by restricting our considerations to events which lie on the x axis in $\mathcal{K}$. By symmetry, these events must also lie on the x' axis in $\mathcal{K}'$. So we want to find the functional relationship between t', x' and t, x.

A neat way of obtaining these results is to consider radar measurements. Let us imagine that there is an observer, Albert, who is stationary at the origin in $\mathcal{K}'$. We imagine that at some event e, he sends out a light pulse in the positive x direction. The pulse is then reflected back towards him at the event A, and received at r. Since $x'_e = x'_r = 0$, and we know that the light pulse always travels at c in $\mathcal{K}'$, we must have

$$x'_A = c(t'_A - t'_e) = c(t'_r - t'_A)\,. \tag{2.3}$$

Albert can therefore infer the coordinates of A by reading the times t'_e and t'_r on their clock, and substituting them into the equations

$$t'_A = \frac{t'_r + t'_e}{2} \qquad \text{and} \qquad x'_A = \frac{ct'_r - ct'_e}{2}\,. \tag{2.4}$$

Now as per our discussion of the 'light clock', we cannot assume that a clock carried by Albert in $\mathcal{K}'$ will tick at the same rate as the stationary clocks in $\mathcal{K}$. If we let γ_v be the factor by which the moving clock's ticks are slowed, and recall that, by definition of the standard configuration, Albert starts his clock at $t = 0$, we find

$$t_e = \gamma_v t'_e \qquad \text{and} \qquad t_r = \gamma_v t'_r\,. \tag{2.5}$$

However, we also know that in $\mathcal{K}$ Albert's world line is given by $x = vt$, which we can use together with the fact that the light pulses have speed c to determine these times in terms of t_A and x_A:

$$\begin{gathered} t_A - t_e = \frac{x_A - vt_e}{c} \implies t_e = \frac{ct_A - x_A}{c - v} \\ \text{and} \\ t_r - t_A = \frac{x_A - vt_e}{c} \implies t_r = \frac{ct_A + x_A}{c + v}\,. \end{gathered} \tag{2.6}$$

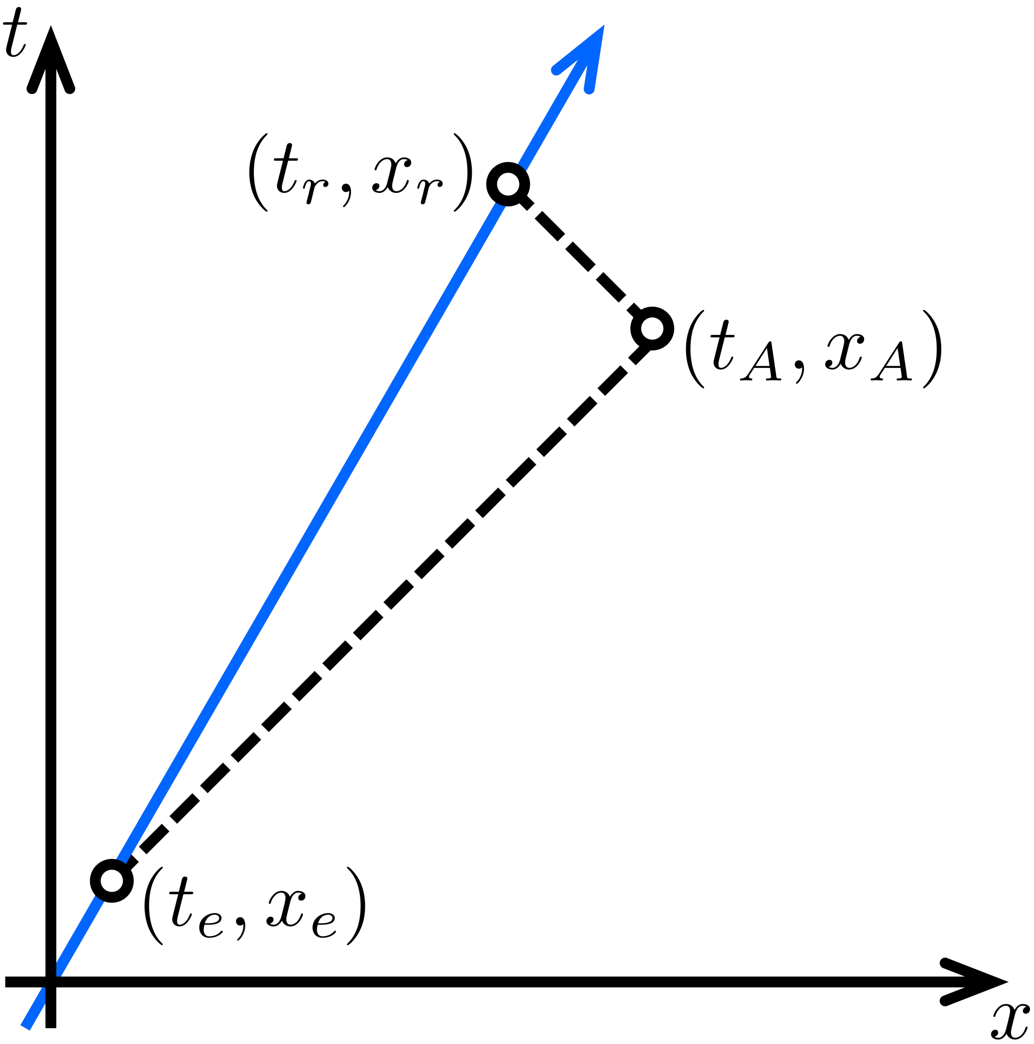


Figure 2. A spacetime diagram of the radar measurement, as viewed from the frame $\mathcal{K}$. Here, and in all subsequent spacetime diagrams, we adopt the convention of plotting in 'spacetime units' (e.g seconds and light seconds) such that light rays are at 45° to the axes. The blue line represents the path followed by the origin of $\mathcal{K}'$, and the dashed lines represent light rays. An observer at the origin of $\mathcal{K}'$ emits a radar pulse at the event e, it is reflected back to them at A, and received at r.

We can combine (2.5) and (2.6) to find the relationship between the coordinates of A in the two frames. This yields

$$t'_A = \frac{t_A - vx_A/c^2}{\gamma_v(1 - v^2/c^2)} \qquad \text{and} \qquad x'_A = \frac{x_A - vt_A}{\gamma_v(1 - v^2/c^2)} \,. \tag{2.7}$$

In our derivation we explicitly assumed that the original radar pulse was emitted in the positive x direction. However, if we repeat the analysis for the negative direction, we will find that the minus signs cancel and we obtain the exact same result. So A can be any event which occurs along the x axis.

Our next task is to determine the form of the function γ_v. To do this we first note that, the un-primed coordinates can be expressed as

$$t_A = \gamma_v \Big(t'_A + v x'_A / c^2 \Big) \qquad \text{and} \qquad x_A = \gamma_v \Big(x'_A + v t'_A \Big) \,. \tag{2.8}$$

We can see that whenever A occurs at the origin in $\mathcal{K}$, it will have $x'_A = -vt'_A$ in $\mathcal{K}'$, so $\mathcal{K}$ must be moving with velocity $-v$ relative to $\mathcal{K}'$. Thus, the transformation from $\mathcal{K}'$ to $\mathcal{K}$ should look the same as (2.7), but with v replaced by $-v$. This can only be true if

$$\gamma_v \gamma_{-v} = \frac{1}{1 - v^2/c^2} \,. \tag{2.9}$$

By symmetry, there should be no difference between the positive and negative directions, so we should have $\gamma_v = \gamma_{-v}$. Using this to solve for γ_v, and choosing the positive root to preserve the forward direction of time, we obtain

$$\gamma_v = \frac{1}{\sqrt{1 - v^2/c^2}} \,. \tag{2.10}$$

This factor appears quite often in special relativity, so it is given a special name: the Lorentz factor. As a general rule of thumb, the percentage error in a calculation carried out using Newtonian physics, instead of relativity, will be approximately the same as the percentage by which γ_v differs from unity. For small velocities this is given roughly by $v^2/2c^2$, so we typically need velocities in excess of $c/10$ for relativity to become noticeable outside of high precision measurements.

We've found the coordinate transformations for events on the x axis, but what about everywhere else? Consider an event B which has coordinates $(t_B, x_B, y_B, 0)$ in $\mathcal{K}$. Let us now imagine all rays of light through B which cross the x axis at some point. By Pythagoras' theorem, these crossings will occur at the events $(t_A, x_A, 0, 0)$ which satisfy

$$c^2 (t_A - t_B)^2 = (x_A - x_B)^2 + y_B^2 \,. \tag{2.11}$$

Using our expressions for t_A and x_A in terms of t'_A and x'_A, we can, after some considerable effort, rearrange this into the form

$$c^2 \left(t'_A - \gamma_v \Big(t_B - v x_B / c^2 \Big) \right)^2 = \left(x'_A - \gamma_v \Big(x_B - v t_B \Big) \right)^2 + y_B^2 \,. \tag{2.12}$$

This equation describes the points $(t'_A, x'_A, 0, 0)$ at which light rays emanating from B cross the x' axis in $\mathcal{K}'$. Since there is nothing to break the symmetry between the positive and negative z direction, we can conclude that $z'_B = 0$. Thus (2.12) must be equivalent to

$$c^2 (t'_A - t'_B)^2 = (x'_A - x'_B)^2 + y'^2_B \,, \tag{2.13}$$

from which we can deduce that

$$t'_B = \gamma_v \Big(t_B - v x_B / c^2 \Big) \,, \quad x'_B = \gamma_v \Big(x_B - v t_B \Big) \,, \quad \text{and} \quad y_B = y'_B \,. \tag{2.14}$$

Since any direction in the y, z plane will be equivalent under rotational symmetry about the x axis, we can finally conclude that the coordinates of an arbitrary event must be related by the Lorentz transformations

$$t' = \gamma_v \Big(t - v x / c^2 \Big) \,, \quad x' = \gamma_v \Big(x - v t \Big) \,, \quad y' = y \,, \quad \text{and} \quad z' = z \,. \tag{2.15}$$

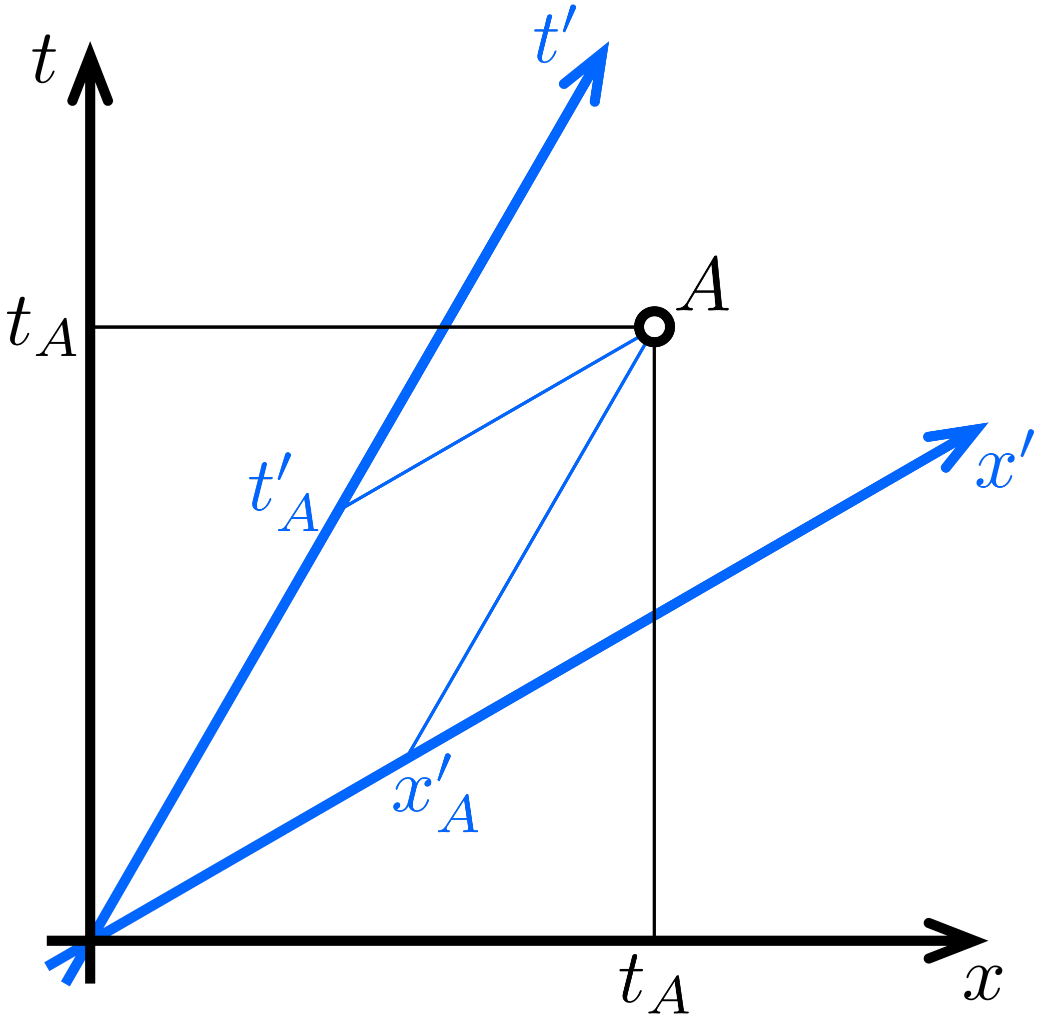


Figure 3. A spacetime diagram showing the relationship between the coordinates of an event A in the inertial frames $\mathcal{K}$ (black) and $\mathcal{K}'$ (blue). Note that instead of being at right angles to one another, the x' and t' axes are oriented such that their angle is bisected by a line at 45°, i.e a light ray.

If we think back to our example of a 'light clock', the invariance of the y and z coordinates is not too surprising. With the light clock we found that the assumption of invariant transverse distances (i.e distances in directions perpendicular to the velocity) was essential for us to obtain the correct rate of ticking for the moving clock, so it should not be too surprising that, once we fixed the ticking rate by symmetry considerations, the invariance of transverse distances was guaranteed.

After a little algebra we can rearrange (2.15) to find expressions for the coordinates of $\mathcal{K}$ in terms of their primed counterparts. This yields

$$t = \gamma_v\Big(t' + vx'/c^2\Big)\,, \quad x = \gamma_v\Big(x' + vt'\Big)\,, \quad y = y'\,, \quad \text{and} \quad z = z'\,, \qquad (2.16)$$

which is simply the same transformation with a relative velocity of $-v$ instead.

2.2 Can you go faster than light?

One of the most immediately apparent facts about the Lorentz transformations is that, if we try to set $v > c$, γ_v becomes imaginary, and the equations become essentially meaningless. It's not too hard to see why. Imagine two light rays propagating along the positive and negative x axes of an inertial frame $\mathcal{K}$. If some other frame $\mathcal{K}'$ has a velocity greater than c along this axis, then in that frame both light rays will be moving in the same direction. However, the two rays cannot have the same speed otherwise they would overlap, which means that at least one must have a speed that is not c. Thus, by the light postulate we can conclude that $\mathcal{K}'$ cannot be an inertial frame.

It might be tempting to interpret this result as a statement that nothing can travel faster than light; however, this is not quite true. What we have actually demonstrated is that no two inertial frames can ever move faster than light relative to one another, but we cannot rule out the possibility that there is some exotic phenomenon in the universe which propagates faster than the speed of light.[1]

That being said, we can make a more general argument if we assume that the universe works according to the basic principles of cause and effect. Let us suppose that an event A is caused by some other event B. A reasonable sounding law of physics is that causes must proceed their effects, and by the principle of relativity we must therefore find that $t_A > t_B$ in any inertial frame. If we suppose that this is true in $\mathcal{K}$, then it can be shown using the Lorentz transformations (2.15) that the condition for it to hold in any other inertial frame $\mathcal{K}'$ is

$$r_{AB} < c(t_A - t_B)\,, \tag{2.17}$$

where r_{AB} is the distance between the events A and B. This result essentially tells us that, if causality is well defined at all, then causal signals cannot travel faster than light. Thus, we find that in special relativity the constant c plays a far more important role than we might have first imagined. It is no longer merely the speed at electromagnetic radiation traverses the vacuum, it now holds pride of place among physical constants as the speed of causality itself.

No discussion of the speed of causality would be complete without at least mentioning quantum entanglement. Suppose we create an electron-positron pair in a quantum superposition of a state where the electron is spin up and the positron is spin down, and a state where the electron is spin down while the positron is spin up. The two particles are separated and sent to two scientists, Emmy and Albert, who have laboratories at the North and South pole respectively. If Emmy now measures her particle she has a 50% chance of finding it to be spin up and a 50% chance of finding it to be spin down. The problem is that, as soon as she makes her measurement the wave function collapses, and Albert's particle changes from being a superposition of up and down to being the opposite of whatever Emmy measured. This problem occurs even if Albert decides to measure his particle in the $\sim 200\,\mathrm{ms}$ it would take for light emitted during Emmy's measurement to reach him. So it

[1] For example, the wave equation $\partial^2\phi/\partial t^2 - c^2\nabla^2\phi - \mu^2\phi = 0$ is completely compatible with special relativity, but for non-zero μ, excitations in ϕ can propagate faster than light.

seems as though Emmy's measurement has had a faster than light causal influence on Albert's particle.

To resolve this issue we need to be slightly more careful with our terminology. Technically speaking, all relativity tells us is that, if two events are too far apart for light to travel between them, then the ordering of those events in time cannot matter. This does not necessarily imply that there can be no causal effects between the two events, it simply means that there can be no transmission of information between them. This is the case here as, while it is true that the outcome of Albert's measurement is determined the moment Emmy makes hers, there is no way he could possibly tell this. Whether or not Emmy makes a measurement at all, he still has a 50% probability of measuring spin up and a 50% probability of measuring spin down. As such, Albert cannot gain any information about whether Emmy made a measurement or not until he waits for a subluminal signal from Emmy.

2.3 Time dilation

While the prediction that causality has a finite speed is certainly significant, it is not too hard to accept. After all, it is not altogether too dissimilar to phenomena in Newtonian mechanics, such as the fact that disturbances in an elastic material can only propagate at the speed of sound. The dual phenomena of time dilation and length contraction are a different story all together. Before we dive into a discussion of these effects, it will be worthwhile for us to introduce a tool which will greatly simplify our calculations. We define the interval $(\Delta s)^2$ between two events as

$$(\Delta s)^2 = (c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 \,, \tag{2.18}$$

where deltas indicate coordinate differences between the two events. The motivation behind this definition is that, if the two events are connected by a light ray, the interval will necessarily be equal to zero by the second postulate. Furthermore, if we calculate the interval between these events in any other inertial frame $\mathcal{K}'$

$$(\Delta s')^2 = (c\Delta t')^2 - (\Delta x')^2 - (\Delta y')^2 - (\Delta z')^2 \,, \tag{2.19}$$

the second postulate guarantees that it must also be zero. In fact, there is an even stronger statement than this. Using the Lorentz transformations (2.15), it is merely a matter of algebra to show that the interval between any two events is invariant, i.e that

$$(\Delta s)^2 = (\Delta s')^2 \,, \tag{2.20}$$

for any two inertial frames $\mathcal{K}$ and $\mathcal{K}'$.[2] While this result does not convey any more information than was already contained within the Lorentz transformations, it can be a very useful way to eliminate some tedious algebra when solving problems.

[2] In fact to show that this is true for any pair of inertial frames requires a little more legwork since we have technically only proved it for the standard configuration. The full proof essentially amounts to noting that translations and rotations preserve $(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2$ without affecting t, and hence must also preserve the interval. We then just need to argue that the conversion between any two inertial frames can be achieved by a combination of translations, rotations, and boosts in the normal configuration.

For example, let us consider the time dilation of a moving clock. By definition, a clock must tick at one second per second in its rest frame, and so by counting the number of ticks (i.e looking at the reading on the clock) we are measuring the time that has elapsed in the clock's rest frame. This time is commonly referred to as the clock's proper time and denoted by τ to differentiate it from coordinate time t. If the clock is inertial, then τ is just the time coordinate t' in its rest frame $\mathcal{K}'$, so the interval between two events on its world line must be

$$(\Delta s')^2 = (c\Delta t')^2 - (\Delta x')^2 - (\Delta y')^2 - (\Delta z')^2 = (c\Delta\tau)^2 \,. \qquad (2.21)$$

Here we have used the fact that $\Delta x' = \Delta y' = \Delta z' = 0$, because any event on the clock's world line must occur at the same place in the clock's rest frame. In some other inertial frame $\mathcal{K}$ the clock moves with speed v. By definition, this means that for two events along its world line

$$v = \frac{\sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2}}{\Delta t} \,, \qquad (2.22)$$

which we can insert into the expression for the interval to obtain

$$(\Delta s)^2 = (c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 = (c^2 - v^2)(\Delta t)^2 \,. \qquad (2.23)$$

Comparing these two expressions for the invariant interval, we find that the moving clock does indeed tick more slowly than it would if it were stationary.

$$\Delta\tau = \sqrt{1 - v^2/c^2}\Delta t = \frac{\Delta t}{\gamma_v} \qquad (2.24)$$

One potentially confusing aspect of this result is that, by the principle of relativity, the clock could equally well claim to be at rest, with any stationary clocks in $\mathcal{K}$ moving relative to it. An identical calculation would lead us to conclude that exactly the same time dilation effect applies, and so both clocks think the other is ticking more slowly than them. The key to understanding this apparent contradiction lies in the relativity of simultaneity. Since the Lorentz transformations express t' as a function of both t and x, it follows that events occurring at the same time in $\mathcal{K}$ do not necessarily occur at the same time in $\mathcal{K}'$. To see how this works, let us consider a simple example. Suppose that at 12 noon a clock flies past you at $\sim 60\%$ the speed of light. From your perspective the moving clock is time dilated by a factor of 0.8, which means that you consider your watch reading 1 p.m to be simultaneous with the clock measuring 48 minutes since it passed you. However, in the clock's frame, its measurement of 48 minutes is considered to be simultaneous with your watch striking 12:38 instead of 1:00. This is shown graphically in Fig. 4. Although the lines of simultaneity of the moving frame are more closely spaced than those of the stationary frame, the 'angle' of the clock's word line through spacetime means that it encounters them less frequently.

So far, everything we have said technically only applies to clocks which are not accelerating. Dealing with accelerations is significantly more complex because not all clocks respond to acceleration in the same way. We could argue that all clocks moving at a given velocity must tick at the same rate using the principle of relativity and the fact that, by definition, two clocks must tick at the same rate when at rest

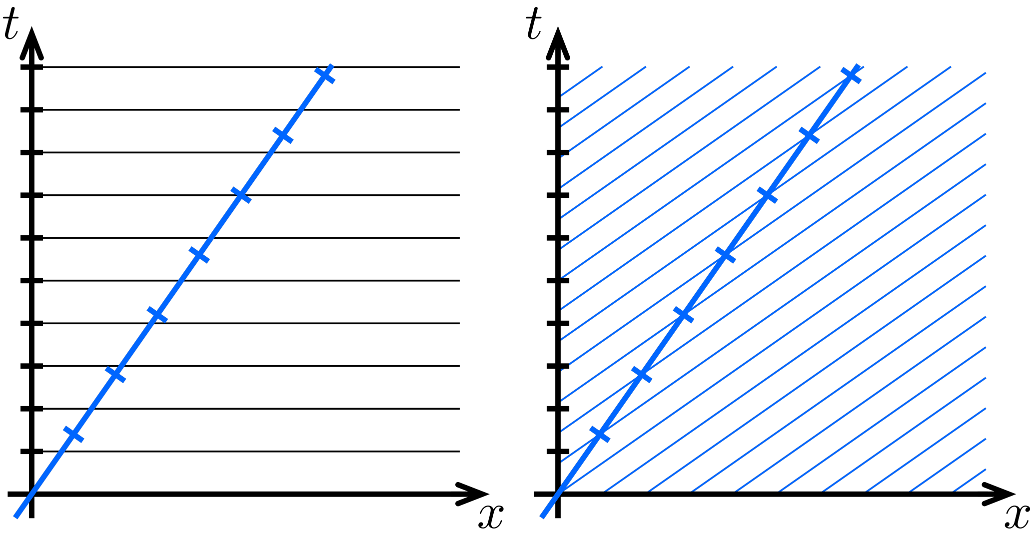


Figure 4. Spacetime diagrams of a clock (blue) moving with a velocity $v = 0.7\,c$ relative to an inertial frame $\mathcal{K}$ (black). The ticks along the clock's world line are separated by the same amount of proper time as the ticks along the time axis. On the left lines of simultaneity of $\mathcal{K}$ are shown, whereas on the right lines of simultaneity for the clock's rest frame $\mathcal{K}'$ are shown.

with respect to one another otherwise they could not both be measuring time. However, the principle of relativity does not make any statements about the relationship between inertial and accelerating frames, so we cannot make any deductions about the relative tick rates of different clocks undergoing the same acceleration. In fact, this idea is really not as novel to us as it might seem. For example, take pretty much any mechanical device, shake (i.e accelerate) it hard enough, and it will eventually cease to function properly. Moreover, the threshold at which it begins to fail, and the exact nature of the failure will depend quite strongly on exactly how the device is constructed. If we are going to take this view that the variation in different clocks' tick rates is a consequence of acceleration induced failure, we should probably have a definition of an ideal clock against which we can compare any real clock to see how well it holds up. One way of approaching this idea is to start by imagining a clock that is subjected to a sequence of instantaneous impulses, interspersed with periods of non-accelerated motion. In general, the clock's reading might jump sharply after each impulse, for example if the sudden force caused a gear to jerk into place too soon; however, it seems pretty reasonable to define an ideal clock as one which exhibits no such jumps. Between two impulses the clock is inertial, and our previous results apply, so we can express the total proper time measured by the clock as

$$\Delta\tau = \sum_i \sqrt{1 - v_i^2/c^2}\,\Delta t_i\,, \tag{2.25}$$

where v_i is the clocks velocity after the ith impulse, and Δt_i is the time between the ith and $(i+1)$th impulses. We can extend this to the general case by defining an ideal clock as one which obeys this relationship no matter how small the time between the impulses gets, and then taking the limit as $\Delta t_i \to 0$. This limit is essentially just a Riemann sum, and so we arrive at the general expression

$$\Delta\tau = \int \sqrt{1 - v^2/c^2}\, dt\,. \tag{2.26}$$

Essentially, we have defined an ideal clock as one which, at any moment, ticks at the same rate as an inertial clock travelling at its instantaneous velocity. So how ideal are real clocks? As a general rule, it is a much easier engineering task to design a clock that will be essentially ideal up to some acceleration threshold than it is to actually achieve that acceleration. This means that, in any practically achievable scenario, it will always be possible to construct a clock whose departures from ideality are negligible. In light of this, we shall always assume that any clock we encounter can be safely approximated as ideal, unless explicitly stated otherwise.

2.4 Length contraction

Looking at the (on axis) Lorentz transformations

$$t' = \gamma_v\Big(t - vx/c^2\Big) \qquad \text{and} \qquad x' = \gamma_v\Big(x - vt\Big)\,, \tag{2.27}$$

we can see that, give or take some factors of c needed for dimensional consistency, time and space enter into them in a symmetric manner. We might then wonder if there is spatial analogue of time dilation, and this is precisely what we find in the phenomenon of length contraction. In brief, we find that moving objects are shorter along their direction of motion, when compared to lengths measured in their rest frames.

One nice way to think about length contraction is as the necessary counterpart to time dilation in order to ensure that two observers agree on their velocities relative to one another (give or take a minus sign). To see how this works, let us consider a specific example. Suppose two train stations are connected by a long straight track of length l_0. If an observer who is at rest relative to the stations measures a train to travel between the platforms in a time T, then they will of course conclude that the train has a speed of $v = l_0/T$ relative to them. From the perspective of an observer inside the train, they are at rest while the outside observer and stations are rushing past them. We know that, due to time dilation, the observer inside the train will only measure a time of T/γ_v between the two platforms passing them. So they will conclude that the speed of the platforms relative to them is $u = \gamma_v l/T$, where l is the distance between the platforms in their frame. These two relative speeds can only agree if the distance between the platforms is length contracted in the train frame, and this length contraction must be precisely $l = l_0/\gamma_v$.

We can look at this phenomenon a bit more generally by considering what happens to a moving ruler. Small objects, like clocks, can be modelled as point particles, and are thus described by world lines; however, a ruler necessarily has a certain length, and so must be described by a world sheet. A ruler's world sheet, such as the one shown in Fig. 5, is the two dimensional surface in spacetime which contains the world lines of all the atoms which make it up. Often, the most important parts of a ruler are its two ends, so if you find a world sheet hard to visualise, it is usually

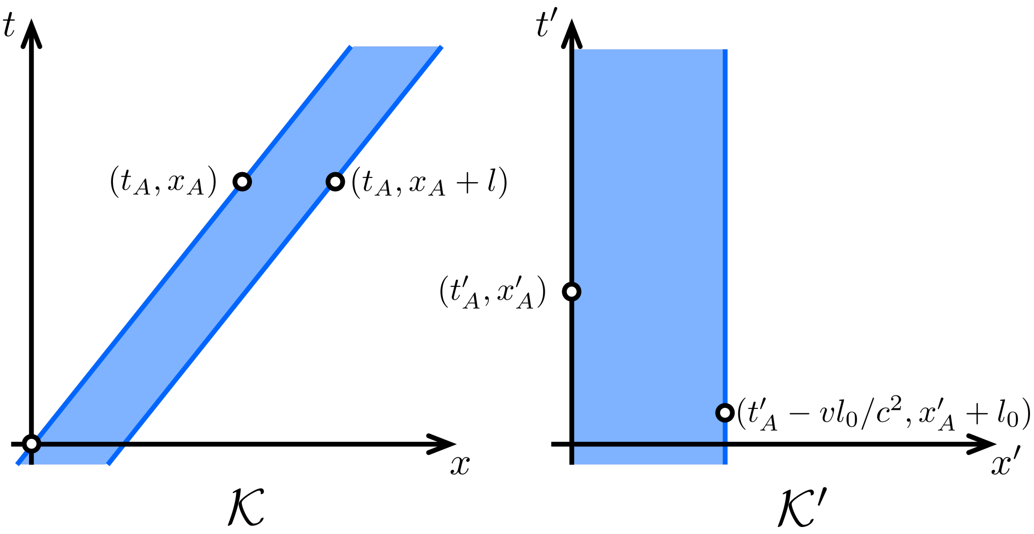


Figure 5. Spacetime diagrams showing the world sheet of a ruler in two different inertial frames. In $\mathcal{K}$ the ruler moves with speed $0.8\,c$, while in $\mathcal{K}'$ it is at rest. Two events which can be used to measure the length of the ruler in $\mathcal{K}$ are shown in each frame.

sufficient to imagine the world lines of the points at each end. A ruler is defined by its rest length, l_0, which is the (constant) distance between its two ends in their mutual rest frame.

Determining the length of a moving ruler is a slightly more complicated issue, since we need to worry about when we measure the positions of the endpoints. The length l of a ruler, as measured in a frame $\mathcal{K}$, is defined to be the distance between two events, one at each end of the ruler, which occur simultaneously in $\mathcal{K}$. This requirement of simultaneity should not come as too much of a surprise; even in Newtonian physics, measuring the distance between a moving ruler's ends at different times wouldn't tell you its length. The main complication special relativity brings is that, since simultaneity is relative, every frame has to use a different pair of events, which would not be necessary in the Newtonian case. We don't have to worry about simultaneity in the ruler's rest frame, because the ends of the ruler are in the same place at all times.

Consider a ruler which moves with a velocity v along the x axis of some inertial frame $\mathcal{K}$. For now, let us assume that the ruler is also aligned along the x axis (i.e parallel to its direction of motion). Let A be an event which occurs at one end of the ruler. The coordinate differences between A and a second simultaneous event on the other end of the ruler are

$$\Delta t = 0, \quad \Delta x = l, \quad \Delta y = 0, \quad \Delta z = 0\,, \tag{2.28}$$

where l is the length of the ruler in $\mathcal{K}$. This is shown in Fig. 5. Thus, we can compute that the interval between the two events is given by

$$(\Delta s)^2 = (c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 = -l^2\,. \tag{2.29}$$

We know that, in the ruler's rest frame $\mathcal{K}'$, the same two events are separated by a distance of l_0 along the x' axis. For the time difference, it is not too hard to see from the Lorentz transformations that simultaneity in $\mathcal{K}$ requires the second event to earlier than A by vl_0/c^2. That is to say that we have

$$\Delta t' = -vl_0/c^2, \quad \Delta x' = l_0, \quad \Delta y' = 0, \quad \Delta z' = 0\,, \tag{2.30}$$

and hence an interval of

$$(\Delta s')^2 = (c\Delta t')^2 - (\Delta x')^2 - (\Delta y')^2 - (\Delta z')^2 = (v^2/c^2 - 1)l_0^2\,. \tag{2.31}$$

We can then use the invariance of the interval to conclude that the two lengths must be related by

$$l = \sqrt{1 - v^2/c^2}\, l_0 = \frac{l_0}{\gamma_v}\,, \tag{2.32}$$

in perfect agreement with what we deduced from our considerations of relative velocities. If we want to consider a ruler which makes an angle to its direction of motion, then we are probably better off dealing directly with the Lorentz transformations. Since the equations are linear, they apply equally as well to the coordinate differences between two events, as they do to the coordinates of the events themselves. So, in general, we have the relations

$$\Delta t' = \gamma_v\Big(\Delta t - v\Delta x/c^2\Big)\,,\ \Delta x' = \gamma_v\Big(\Delta x - v\Delta t\Big)\,,\ \Delta y' = \Delta y\,,\ \text{and}\ \Delta z' = \Delta z\,. \tag{2.33}$$

Specialising these to the case of two events which are simultaneous in $\mathcal{K}$ (i.e $\Delta t = 0$), and rearranging the x equation slightly, we obtain

$$\Delta x = \frac{\Delta x'}{\gamma_v}\,, \quad \Delta y = \Delta y'\,, \quad \text{and} \quad \Delta z = \Delta z'\,. \tag{2.34}$$

So we can see that length contraction only occurs in the direction parallel to the ruler's motion. One important fact to note is that length contraction does not just alter the length of the ruler. Since it only affects lengths in one direction, it also changes the angle that the ruler makes to the x axis.

Just as with time dilation, there is an apparent contradiction that two identical rulers moving relative to one another will both think that the other is shorter than them. The resolution is much the same; since measurements of length depend so heavily on simultaneity in their definition, the differences in simultaneity between the two frames perfectly explains how they can disagree about which of them is longer. Dealing with accelerating rulers turns out to be much more complicated than accelerating clocks. One of the reasons why is that relativity forbids any object from being perfectly rigid (such an object would have a superluminal speed of sound, which would break causality), and so any ruler compatible with special relativity must change shape when a force is applied to it. Another issue is that measuring lengths requires a definition of when two events are simultaneous, which is somewhat non-trivial to establish for an accelerated observer. Due to these conceptual complications, we shall not discuss accelerating rulers in any great detail.

2.5 Velocity addition and rapidity

Suppose that a particle has a velocity (u'_x, u'_y, u'_z) with respect to an inertial frame $\mathcal{K}'$, which is itself moving with a velocity v along the x axis of a frame $\mathcal{K}$. We know that the particle's velocity relative to $\mathcal{K}$ can't simply be $(u'x + v, u'_y, u'_z)$, because if the particle was a photon this would not leave the speed of light invariant. Conceptually, this breakdown in Newtonian logic can be attributed to the time dilation and length contraction between the two frames altering the way in which velocities add together. We can calculate the correct velocity addition formula as follows. First, we note that, since the Lorentz transformations are linear, we can always translate our axes so that the particle passes through the origin at time zero. In this case its world line is given by

$$x' = u'_x t' \,, \qquad y' = u'_y t' \,, \qquad \text{and} \qquad z' = u'_z t' \,. \tag{2.35}$$

We can now use the Lorentz transformations to find out what this world line looks like in the un-primed coordinates. Dealing first of all with the x' equation, we find that

$$\gamma_v \Big(x - vt \Big) = u'_x \gamma_v \Big(t - vx/c^2 \Big) \implies x = \frac{u'_x + v}{1 + vu'_x/c^2} \, t \,. \tag{2.36}$$

Substituting this result into the Lorentz transformed y' equation yields

$$y = u'_y \gamma_v \Big(t - vx/c^2 \Big) = u'_y \gamma_v \left(1 - \frac{vu'_x + v^2}{c^2 + vu'_x} \right) t \,. \tag{2.37}$$

The z' equation will transform in essentially the same way, so after tidying this up[3] we obtain

$$y = \frac{u'_y}{\gamma_v (1 + vu'_x/c^2)} \, t \,, \qquad \text{and} \qquad z = \frac{u'_z}{\gamma_v (1 + vu'_x/c^2)} \, t \,. \tag{2.38}$$

By inspecting these equations and comparing them to (2.35), we can see that they describe the world line of a particle whose velocity in $\mathcal{K}$ is given by

$$(u_x, u_y, u_z) = \left(\frac{u'_x + v}{1 + vu'_x/c^2}, \frac{u'_y}{\gamma_v (1 + vu'_x/c^2)}, \frac{u'_z}{\gamma_v (1 + vu'_x/c^2)} \right) \,. \tag{2.39}$$

This expression is rather complicated and there is not a lot of insight that can be gleaned from looking at it. One important thing to be aware of here is that this formula involves coupling between longitudinal and transverse velocities (i.e u_y and u_z depend on u'_x), which will be relevant when we come to explaining the aberration of starlight.

There are some interesting ideas that we can explore if we consider the special case where the two velocities are collinear (i.e $u_y = u_z = 0$). Dropping the subscript x to avoid clutter, this restriction reduces (2.39) down to give

$$u = \frac{u' + v}{1 + u'v/c^2} \,. \tag{2.40}$$

[3]Remember that $\gamma_v = 1/\sqrt{1 - v^2/c^2}$ so, when we get a factor of $\gamma_v(1 - v^2/c^2)$, it can be simplified to $1/\gamma_v$.

While this result is certainly less ugly than the full formula, it is still a little awkward to deal with. Fortunately for us, there is a way to make it even simpler. If you have done a lot of hyperbolic trigonometry recently, you might notice that introducing the rapidities ξ, ξ', χ defined by

$$u = c \tanh \xi \,, \qquad u' = c \tanh \xi' \,, \qquad \text{and} \qquad v = c \tanh \chi \,, \tag{2.41}$$

allows (2.40) to be written in the form of a standard trigonometric identity,

$$\tanh \xi = \frac{\tanh \xi' + \tanh \chi}{1 + \tanh \xi' \tanh \chi} = \tanh(\xi' + \chi) \implies \xi = \xi' + \chi \,. \tag{2.42}$$

So it turns out that, at least when everything is collinear, rapidities transform in the same way as velocities in Newtonian mechanics, which is to say they add together linearly.[4] The rapidity also gives us a nice way of understanding why it will never be possible to exceed the speed of light by combining relative velocities. No matter how many rapidities we add together, and no matter how large they are, the final value of ξ will still be finite, so $\tanh \xi < 1$, and the particle's velocity will still be subluminal.

It turns out that a lot of the ideas we have already discussed can be fairly nicely expressed in terms of rapidity. For example, the Lorentz factor γ_v, which was so important for time dilation and length contraction, can be expressed as

$$\gamma_v = \frac{1}{\sqrt{1 - v^2/c^2}} = \frac{1}{\sqrt{1 - \tanh^2 \chi}} = \frac{\cosh \chi}{\sqrt{\cosh^2 \chi - \sinh^2 \chi}} = \cosh \chi \,. \tag{2.43}$$

Using this result, and a little bit of algebra to juggle powers of c into the right places, it is possible to write the Lorentz transformations in the elegant form

$$ct' = ct \cosh \chi - x \sinh \chi \,, \;\; x' = x \cosh \chi - ct \sinh \chi \,, \;\; y' = y \,, \;\; \text{and} \;\; z' = z \,. \tag{2.44}$$

In much the same way as the identity $\cos^2 \theta + \sin^2 \theta = 1$ guarantees that Euclidean distance (i.e $x^2+y^2+z^2$) will be unchanged by rotations, (2.44) ensures the invariance of the interval (i.e $c^2t^2 - x^2 - \ldots$) via the corresponding hyperbolic result that $\cosh^2 \chi - \sinh^2 \chi = 1$.

2.6 Waves and the Doppler effect

Wave phenomena are essentially ubiquitous in physics, and so we would be remiss if we did not take some time to examine how special relativity affects the theory of waves. In general terms, an observer in some reference frame $\mathcal{K}$ can describe a wave with a function $\psi(t, x, y, z)$, which assigns a numerical value to every time and place. The physical interpretation of this function depends on the particular type of wave we are considering. It could represent air pressure if we were talking about sound waves, the height of the water level if we were talking about water waves, or the

[4] We might wonder if this idea could be extended beyond the case of collinear velocities. Unfortunately, this doesn't really work, because in (2.39) the two velocities do not commute, so it cannot be represented by a commutative process like addition.

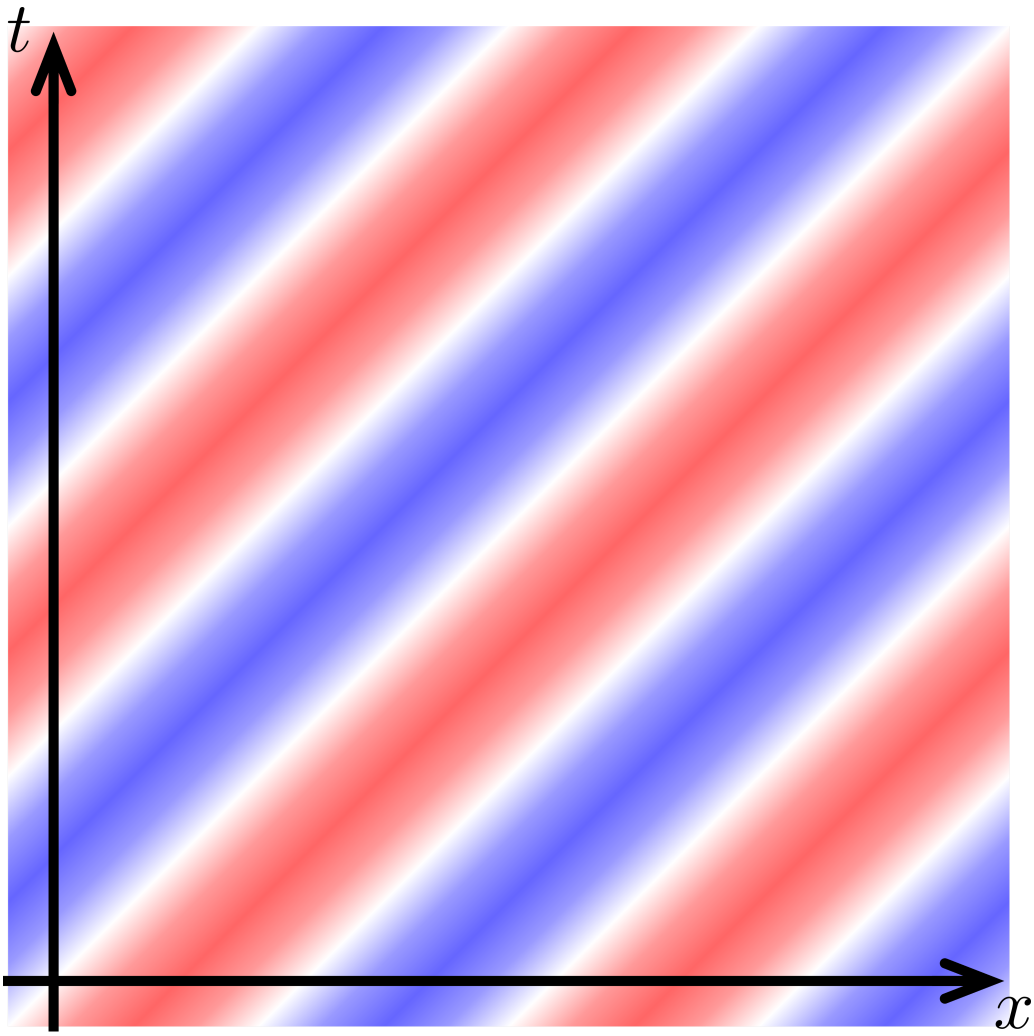


Figure 6. A visual depiction of a wave $\psi(t,x)$ in spacetime. The magnitude of ψ is represented by the density of the colour, with red and blue corresponding to positive and negative values of ψ respectively. This diagram shows a wave travelling at the speed of light from left to right. This is shown by the fact that the contours of constant colour are lines at 45° to the spacetime axes.

electric and magnetic field strengths if we were talking about light.[5] If the frequency and wavelength of a wave according to an observer in $\mathcal{K}$ are ν and λ respectively, then its function will take the general form

$$\psi(t,x,y,z) = A\cos\left(2\pi\nu t - \frac{2\pi}{\lambda}(x\cos\theta_x + y\cos\theta_y + z\cos\theta_z) + \phi\right) . \tag{2.45}$$

The constants A and ϕ denote the amplitude and phase of the wave, while $\theta_x, \theta_y, \theta_z$ are the angles between the wave's direction of propagation and each of the coordinate axes respectively. To see that the parameters ν and λ actually correspond to the

[5]Technically speaking, in the case of light we should think of ψ not as a single number but rather as a list of six different numbers representing the components of the electric and magnetic field vectors. The analysis in this case is slightly more complicated but the general conclusions are the same so we shall not concern ourselves with this subtlety.

physical frequency and wavelength of the wave we will need to make use of the fact that the cosine function is periodic with period 2π. Thus, if we consider two events at the same location in space but separated in time by $\Delta t = 1/\nu$, then the arguments of the appropriate cosines will differ by 2π and we will have

$$\psi(t + 1/\nu, x, y, z) = \psi(t, x, y, z) \,. \tag{2.46}$$

In other words, at a particular location the wave oscillates with a time period of $1/\nu$, which implies that its frequency is indeed ν. A similar argument can be applied to the wavelength. In this case we consider two events which occur at the same time, but separated by a distance λ in the direction of the wave's propagation. Drawing three separate right angled triangles, each taking the line joining the two events as its hypotenuse and with another side parallel to one of the coordinate axes, simple trigonometry tells us that the differences in coordinates between the two events are

$$(\Delta x, \Delta y, \Delta z) = (\lambda \cos\theta_x, \lambda\cos\theta_y, \lambda\cos\theta_z) \,. \tag{2.47}$$

Arguments from Euclidean geometry, which essentially boil down to an application of Pythagoras' theorem, tell us that the direction cosines are not all independent from one another and must satisfy the relation

$$\cos^2\theta_x + \cos^2\theta_y + \cos^2\theta_z = 1 \,. \tag{2.48}$$

Thus the difference in the argument of the cosine in (2.45) for the two events will once again be equal to 2π, and as such we will have

$$\psi(t, x + \lambda\cos\theta_x, y + \lambda\cos\theta_y, z + \lambda\cos\theta_z) = \psi(t, x, y, z) \,, \tag{2.49}$$

confirming that the wavelength of the wave described by $\psi(t, x, y, z)$ is in fact λ. In addition to the frequency and wavelength, we can also determine the speed at which a wave propagates. We say that a wave has phase speed V, if the value of ψ is constant along any worldline which would correspond to a particle moving at speed V along the direction specified by the angles $\theta_x, \theta_y, \theta_z$. That is to say that the equation

$$\psi(t + \Delta t, x + V\Delta t\cos\theta_x, y + V\Delta t\cos\theta_y, z + V\Delta t\cos\theta_z) = \psi(t, x, y, z) \,, \tag{2.50}$$

should hold true for all values of Δt. Substituting in the explicit expression for $\psi(t, x, y, z)$ given in (2.45), we find that the phase speed V must satisfy

$$2\pi\nu - \frac{2\pi}{\lambda}(V\cos^2\theta_x + V\cos^2\theta_y + V\cos^2\theta_z) = 0 \,. \tag{2.51}$$

Using relation (2.48) between the direction cosines, rearranging, and simplifying yields the well known formula

$$V = \nu\lambda \,. \tag{2.52}$$

So far, we have centred our entire discussion upon a single inertial frame of reference; however, it should not be surprising that what we are most interested in is really how the wave properties change between different reference frames. Since it holds true for the majority of real physical systems, we shall assume that the quantity ψ obeys a simple transformation law of the form

$$\psi'(t', x', y', z') = L\psi(t, x, y, z) \,, \tag{2.53}$$

where L is some function of the relative velocity between the frames $\mathcal{K}$ and $\mathcal{K}'$.[6] The key point is that the wave is local; the value of ψ' measured by an observer in $\mathcal{K}'$ at a particular event depends linearly upon the value of ψ measured by an observer in $\mathcal{K}$ at that same event. Substituting in (2.45) for $\psi(t,x,y,z)$, we find that

$$\psi'(t',x',y',z') = A'\cos\left(2\pi\nu t - \frac{2\pi}{\lambda}(x\cos\theta_x + y\cos\theta_y + z\cos\theta_z) + \phi\right), \tag{2.54}$$

where $A' = LA$ is the wave's amplitude measured by an observer in $\mathcal{K}'$. If we want to determine this observer's perception of the frequency and wavelength, we will need to know how ψ' varies as a function of the primed coordinates t', x', y', z'. To this end, we can substitute in the inverse Lorentz transformations (2.16) to obtain

$$\begin{aligned}\psi'(t',x',y',z') = A'\cos\Bigg(2\pi\gamma_v\Big(\nu - \frac{v\cos\theta_x}{\lambda}\Big)t' - \frac{2\pi\gamma_v x'}{\lambda}\big(\cos\theta_x - v\nu\lambda/c^2\big)\\ - \frac{2\pi y'}{\lambda}\cos\theta_y - \frac{2\pi z'}{\lambda}\cos\theta_z + \phi\Bigg).\end{aligned} \tag{2.55}$$

We know that a wave with frequency and wavelength ν' and λ' which propagates in the direction specified by the angles $\theta'_x, \theta'_y, \theta'_z$ in the frame $\mathcal{K}'$ would be described by a function of the form

$$\psi'(t',x',y',z') = A'\cos\left(2\pi\nu' t' - \frac{2\pi}{\lambda'}(x'\cos\theta'_x + y'\cos\theta'_y + z\cos\theta'_z) + \phi'\right). \tag{2.56}$$

By comparing these two expressions, we can immediately see that the phase constants ϕ and ϕ' for the two frames must be equal. This is a direct consequence of the fact that we have chosen our frames such that their origins coincide at time zero. Somewhat more interestingly, we can see that the frequency measured by an observer in $\mathcal{K}'$ is related to that measured by an observer in $\mathcal{K}$ by

$$\nu' = \gamma_v\big(\nu - v\cos\theta_x/\lambda\big) = \gamma_v\big(1 - v\cos\theta_x/V\big)\nu\,, \tag{2.57}$$

where $V = \nu\lambda$ is the phase speed of the wave in $\mathcal{K}$. While this formula specifically only applies to two frames in the standard configuration, we can generalise it to any pair of frames by replacing θ_x with the angle between the velocity of $\mathcal{K}'$ and the wave's propagation direction, as measured in $\mathcal{K}$. This can be seen by noting that we can transform between any pair of frames by first rotating our coordinate axes so that they are in the standard configuration, then applying the appropriate Lorentz transformations, and finally rotating the axes back into position, together with the fact that angles do not change with rotations of the coordinate axes.

A particularly common physical situation to encounter is to find that there exists some inertial frame $\mathcal{K}$ in which the waves we are interested in travel at a known phase speed V. We now consider waves which are emitted by some specific source and received by a particular observer, both of which may be in motion relative to $\mathcal{K}$. We can use (2.57) to relate the frequency ν_{em} in the source frame, which is usually

[6] For something like an electromagnetic wave where ψ should be interpreted as a list of different components, then L should be interpreted as a matrix which mixes those components with one another. This leads to some interesting effects regarding how the polarisation of the wave changes between reference frames, but does not affect any of the central conclusions we draw here.

fixed by some internal property of the source, to the frequency ν in $\mathcal{K}$. We can then apply (2.57) once again to relate this frequency to the frequency ν_{rec} measured by the observer. Combining these relations yields

$$\frac{\nu_{\text{rec}}}{\nu_{\text{em}}} = \frac{\gamma_{u_{\text{rec}}}(V - u_{\text{rec}}\cos\theta_{\text{rec}})}{\gamma_{u_{\text{em}}}(V - u_{\text{em}}\cos\theta_{\text{em}})} = \sqrt{\frac{c^2 - u_{\text{em}}^2}{c^2 - u_{\text{rec}}^2}}\,\frac{V - u_{\text{rec}}\cos\theta_{\text{rec}}}{V - u_{\text{em}}\cos\theta_{\text{em}}}\,. \tag{2.58}$$

Here u_{em} and u_{rec} are the speeds of the source and observer respectively, while θ_{em} and θ_{rec} are the angles between their velocities and the wave's direction of propagation, all measured in the frame $\mathcal{K}$.

The dependence of the observed frequency on the motions of the source and receiver is known as the Doppler effect. To get a feeling for how this works in practice, we can look at an example that should be familiar from our day to day lives. When a car drives past you, the sound from its engine seems to change in pitch, being higher while its approaching you and lower as it drives away. Sound waves propagate with a known phase velocity $V \approx 340\,\mathrm{m\,s^{-1}}$ in a frame where the air they travel through is at rest. Human walking speed is slow compared to a car, and so we can approximate $u_{\text{rec}} \approx 0$. For a reasonably fast moving road, a typical speed for a car would be $u_{\text{em}} \approx 20\,\mathrm{m\,s^{-1}}$. The value of $\cos\theta_{\text{em}}$ will change from $+1$ while the car is driving towards you, to -1 once it's heading away from you. Using these numbers, and noting that the frequency of the engine noise in its own frame ν_{em} will be roughly constant, equation (2.58) tells us that the sound frequency will be reduced by a factor of approximately 8/9 after the car has passed you.[7] In terms of music theory, this frequency ratio corresponds to a lowering of the pitch by one major tone.

One important thing to note about the Doppler effect is that it doesn't just apply to the frequency of the wave itself, but also to the frequency of any modulations in the wave's amplitude. This is illustrated in Fig. 7, which shows the effect of a Doppler shift on a pulsed waveform. One way of understanding this phenomenon is to recognise that an amplitude modulated wave can be mathematically re-expressed as a linear combination of pure frequency waves. For a simple example, we can consider the identity

$$\cos(2\pi\nu_{\text{mod}}t)\cos(2\pi\nu t) = \frac{\cos(2\pi(\nu + \nu_{\text{mod}})t) + \cos(2\pi(\nu - \nu_{\text{mod}})t)}{2}\,. \tag{2.59}$$

The left hand side represents a wave (at some specific point in space) with a frequency ν, whose amplitude is oscillating at some second frequency ν_{mod}. This is exactly equivalent to the right hand side, which represents a superposition of two sinusoidal waves, one with frequency $\nu + \nu_{\text{mod}}$ and one with $\nu - \nu_{\text{mod}}$. When transforming into a different reference frame, both of the waves on the right hand side are of the form we have been considering, and so their frequencies will change according to the Doppler effect formula (2.58). This then implies that, on the left hand side, both the carrier frequency ν and the modulation frequency ν_{mod} will be affected.

[7] Since all the speeds in this example are much slower than light, the factor of $\sqrt{(c^2 - u_{\text{em}}^2)/(c^2 - u_{\text{rec}}^2)} \approx 1$ and can be neglected. This factor contains the relativistic corrections to the result which would be obtained in Newtonian physics, and so this is telling us that special relativity was not really necessary in this example, which we probably could have guessed.

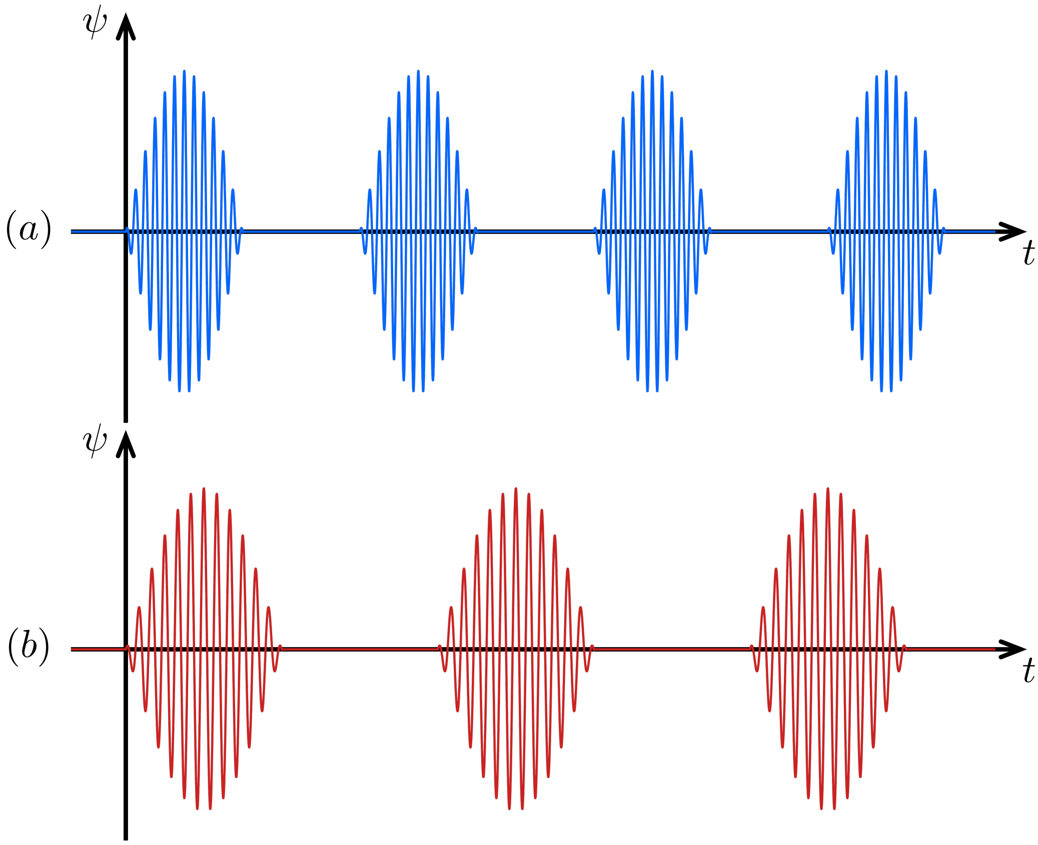


Figure 7. The waveform on the top corresponds to some source of reasonably well defined frequency (e.g a laser pointer) being switched on and off. Below is that same waveform but after a slight Doppler shift. Note that not only is the frequency within the pulses lower, but each pulse is now broader, and the time between them is also longer, all by the same proportionality factor.

Observers in different frames of reference will also measure different values for the wavelength of a particular wave. We can analyse this phenomenon in much the same way as the Doppler effect. First, we can compare (2.55) and (2.56) to deduce that we must have

$$\frac{\cos\theta'_x}{\lambda'} = \frac{\gamma_v\big(\cos\theta_x - v\nu\lambda/c^2\big)}{\lambda}\,, \quad \frac{\cos\theta'_y}{\lambda'} = \frac{\cos\theta_y}{\lambda}\,, \quad \text{and} \quad \frac{\cos\theta'_z}{\lambda'} = \frac{\cos\theta_z}{\lambda}\,. \tag{2.60}$$

In order to extract the wavelength λ' from these relations, we can invoke (2.48) applied in the frame $\mathcal{K}'$ to argue that the squares of the direction cosines should sum to one. Thus, we can deduce that λ' is given by

$$\lambda' = \frac{1}{\sqrt{\left(\frac{\cos\theta'_x}{\lambda'}\right)^2 + \left(\frac{\cos\theta'_y}{\lambda'}\right)^2 + \left(\frac{\cos\theta'_z}{\lambda'}\right)^2}}\,. \tag{2.61}$$

Substituting in the expressions for the primed quantities in terms of their unprimed counterparts, using (2.48) to simplify slightly, and substituting in $V = \nu\lambda$ leads us to the wavelength transformation law

$$\lambda' = \frac{\lambda}{\sqrt{1 + (\gamma_v^2 - 1)\cos^2\theta_x - 2\gamma_v^2 vV\cos\theta_x/c^2 + \gamma_v^2 v^2 V^2/c^4}}\,. \tag{2.62}$$

Just as before, we can generalise this result to the case when $\mathcal{K}$ and $\mathcal{K}'$ are not in the standard configuration by replacing θ_x with the angle between the frames' relative velocity and the wave's direction of travel. Furthermore, we can essentially repeat the same arguments we used before to describe the situation where we have a source and an observer moving relative to a frame in which the wave's phase speed is known. In this case we find that the emitted and received wavelengths are related by

$$\frac{\lambda_{\text{rec}}}{\lambda_{\text{em}}} = \sqrt{\frac{c^4 + (\gamma_{u_{\text{em}}}^2 - 1)c^4 \cos^2\theta_{\text{em}} - 2\gamma_{u_{\text{em}}}^2 u_{\text{em}} V c^2 \cos\theta_{\text{em}} + \gamma_{u_{\text{em}}}^2 u_{\text{em}}^2 V^2}{c^4 + (\gamma_{u_{\text{rec}}}^2 - 1)c^4 \cos^2\theta_{\text{rec}} - 2\gamma_{u_{\text{rec}}}^2 u_{\text{rec}} V c^2 \cos\theta_{\text{rec}} + \gamma_{u_{\text{rec}}}^2 u_{\text{rec}}^2 V^2}} . \tag{2.63}$$

The general formulae (2.58) and (2.63) are quite cumbersome and difficult to deal with, so it seems sensible to look for some special cases in which they can be simplified. A particularly useful case to consider arises when the waves propagate with a phase speed $V = c$, which would be the case for light in a vacuum. The light postulate ensures that, if this is true in one frame, it will hold true in all frames, so we are free to analyse the situation from the rest frame of the observer. It can then be shown after a little algebra that, in this case, the general formulae reduce to the simpler form

$$\frac{\nu_{\text{rec}}}{\nu_{\text{em}}} = \frac{\lambda_{\text{em}}}{\lambda_{\text{rec}}} = \frac{\sqrt{c^2 - v^2}}{c - v\cos\theta} , \tag{2.64}$$

where v is the speed of the source in the observer's rest frame, and θ is the angle between its velocity and the wave's direction of travel. We can simplify this result even further if we assume that the source is moving directly towards the observer.[8] In this case the source will be moving parallel to the wave, and so we can set $\theta = 0 \implies \cos\theta = 1$, and noting after that $c^2 - v^2 = (c+v)(c-v)$ this yields

$$\frac{\nu_{\text{rec}}}{\nu_{\text{em}}} = \frac{\lambda_{\text{em}}}{\lambda_{\text{rec}}} = \sqrt{\frac{c+v}{c-v}} = e^{\chi} , \tag{2.65}$$

where χ is the rapidity of the source, i.e $\tanh\chi = v/c$. When talking about light signals it is common to use the terms redshift and blueshift to denote the direction of the Doppler effect. If light is received with a higher frequency than it was emitted with, which would happen if the source was moving towards the observer, then we describe it as being blueshifted, because its frequency has moved towards the blue end of the visible spectrum. Conversely, if the light is detected at a lower frequency, which would correspond to a source moving away from the observer, we say it has been redshifted, since its frequency moved towards the red end of the spectrum.

[8] We could also obtain the corresponding result for a source moving directly away from the observer, but this simply amounts to replacing v with $-v$.

3 Paradoxes in Special Relativity

One of the unfortunate drawbacks of being human is that we spend our entire lives almost exclusively interacting with objects moving very slowly relative to us. As such, we have essentially zero first hand experience of relativistic phenomena in action. This problem is that our physical intuition, which is derived solely from our experiences, often conflicts with the predictions of special relativity. In order to prepare ourselves for some of the most common ways our dodgy intuition can slip in and spoil our thinking, we shall examine several so called paradoxes which arise in the special theory of relativity.

A true paradox in special relativity would be indicative of a logical contradiction inherent in the theory, and would undermine it on a fundamental level. As evidenced by the fact people still believe in special relativity over a hundred years after its original formulation, such a paradox does not exist. The 'paradoxes' we shall be discussing are of a weaker variety. The apparently paradoxical results stem from misunderstandings in attempting to reconcile special relativity with an intuitive Newtonian view of space and time. Understanding how these misunderstandings arise and how they are resolved will be essential for avoiding similar pitfalls in the future.

3.1 The twins paradox

Perhaps the most famous example we shall examine is the twins paradox. The set-up is as follows. We imagine two twins, Albert and Emmy, who have lived their entire lives on Earth without ever moving at relativistic speeds relative to one another. One day, Emmy is selected to take part in a test flight for a newly constructed interstellar spaceship. She gets into the ship and flies at a speed of eighty percent the speed of light to a star four light-years away, before turning around and returning to Earth at the same speed. The question is: how do the ages of the two twins compare once they are finally reunited?

We can calculate the time taken for the journey in the Earth's frame using elementary kinematics. Emmy has to travel a total distance of 8 lyr, and she will do so at a constant speed of $4c/5 = 0.8\,\mathrm{lyr\,yr}^{-1}$. Thus, in the Earth's frame, Emmy's journey must last 10 yr. Since Albert is always stationary on Earth, this must be the proper time he experiences and so he will have aged 10 yr between Emmy's departure and her return. On the other hand, Emmy will spent the entire journey time dilated by a factor of $\sqrt{1-0.8^2} = 0.6$, so her proper time should be only 6 yr. As such, we Albert will be 4 yr older than Emmy when she returns to Earth. This is indeed the correct answer, and it didn't require an awful lot of effort to obtain, so where exactly is the paradox?

The potential for confusion arises when we consider the same situation from Emmy's perspective. In her frame the Earth and star are both moving at speed $0.8\,\mathrm{lyr\,yr}^{-1}$, and so the distance between them is length contracted to only 2.4 lyr. This makes sense as it will take three years to cover this distance at $0.8\,\mathrm{lyr\,yr}^{-1}$, so Emmy agrees

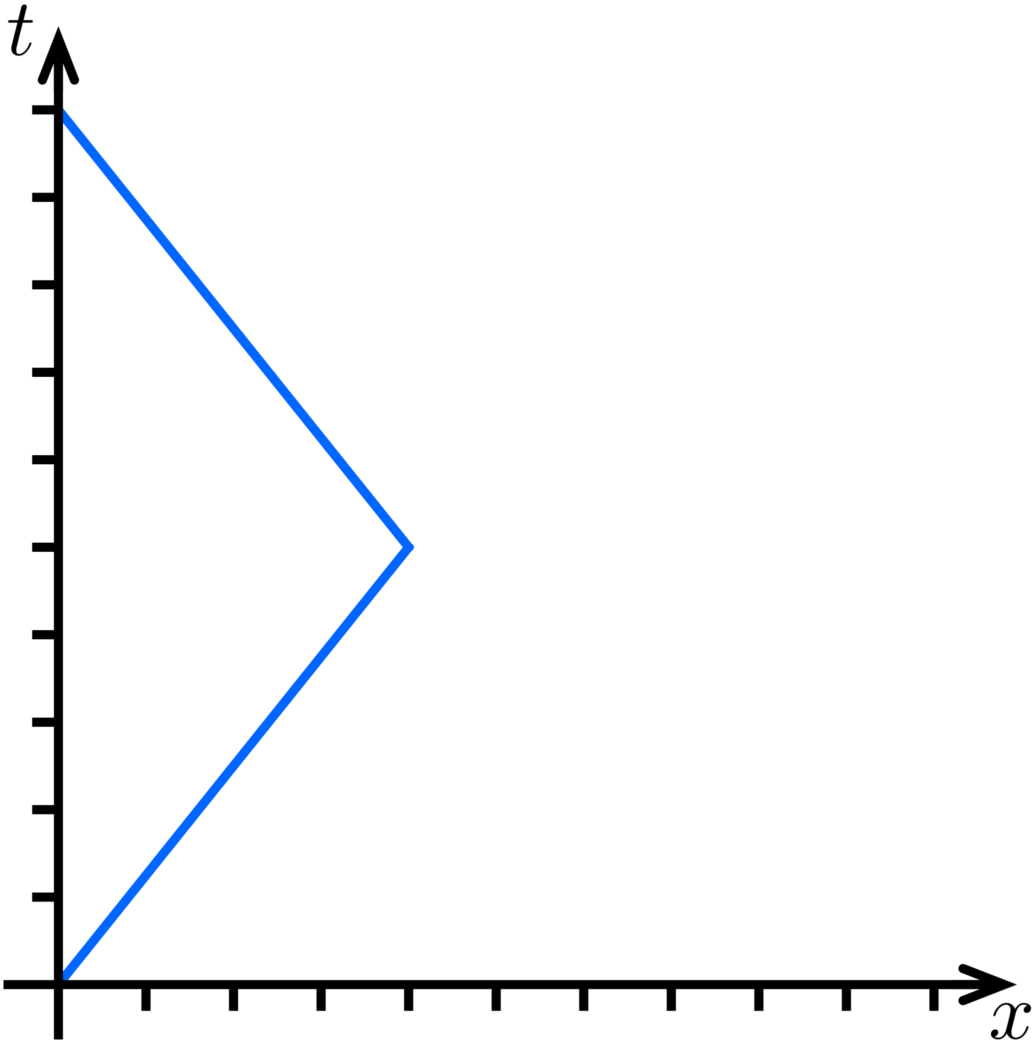


Figure 8. Spacetime diagram of the twins paradox in the Earth's reference frame. Albert's world line lies along the time axis and Emmy's is shown in blue. The divisions marked along the axes are years and light-years respectively.

that her round trip will be 6 yr long. However, in her frame Albert is always moving at speed $4c/5$ and so it is tempting to argue that he should be time dilated by a factor of 0.6, experiencing a total time of only 3.6 yr.

The flaw in this argument lies in the fact that Emmy does not spend her entire journey in a single inertial frame of reference. Instead, when she turns around she switches from a frame moving away from the Earth into one with a velocity towards the Earth. Equivalently, we can say that Emmy's reference frame is accelerating or non-inertial. Since the principle of relativity only enforces the equivalence of different inertial frames, there is no reason why we should assume that applying the formulae we derived about inertial frames to a non-inertial one will give the correct answer. Indeed, as the twins paradox demonstrates, if you pretend that Emmy's frame is inertial and carry out your calculations accordingly, you will get a drastically wrong result.

Still, Emmy is inertial for nearly her entire journey, with the exception of one brief period of acceleration when she turns around[1] and so it is not necessarily obvious how this observation solves all our problems. After all, up until the moment before she turns around, she is inertial and so it is true that Albert is time dilated relative to her, having only measured a time of 1.8 yr compared to her 3 yr. The same is true for the return journey from the moment just after she turns around until her return. So it seems as though the entirety of Albert's additional 6.4 yr must be accounted for during the brief period of acceleration. In other words, Albert must age infinitely quickly in Emmy's frame, while she is accelerating. This is shown graphically in Fig. 9.

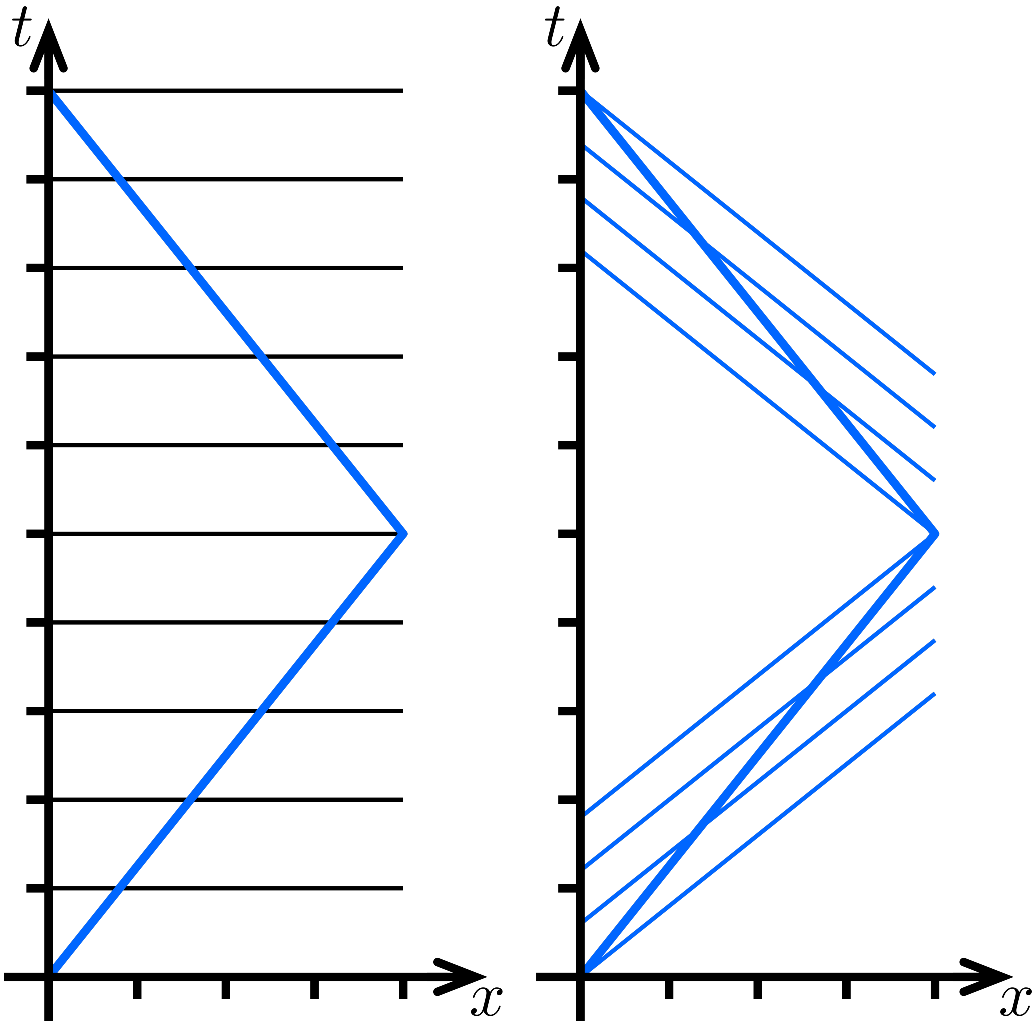


Figure 9. Spacetime diagrams of the twins paradox in the Earth's reference frame. On the left Albert's planes of simultaneity are shown as thin black lines, and on the right Emmy's are shown as thin blue lines. Note that when Emmy turns around her plane of simultaneity rapidly sweeps across 6.4 yr of Albert's time.

[1] Technically, she also accelerates at the moment she leaves Earth and the moment she returns, but these accelerations are not important and can be removed by rephrasing the paradox slightly.

Now we can always argue that, since we are assuming Emmy turns around instantly, her acceleration is infinite, and so it's not exactly surprising that Albert has to age infinitely quickly in her frame. However, a similar problem can emerge even if Emmy's acceleration is finite. For example, let's suppose that it takes Emmy a full year, as measured from Earth, to turn around. She would now need to account for 7.4 yr of Albert's time in the less than 1 yr she spends turning around. While this isn't a huge factor it is larger than we might expect given that her acceleration lies well within the range of what would be achievable on Earth. Furthermore, we can make the effect of the acceleration even more pronounced if we have Emmy travel further before turning around. If she travelled a distance of 30 lyr, then she would have to account for 65 yr of Albert's time in the same less than 1 yr she spent accelerating. In summary, even a small acceleration can have massive consequences if it is far enough away. We can see this in Fig. 9 by noting that the effect of accelerating is to rotate Emmy's plane of simultaneity on the spacetime diagram. It follows that, the further the centre of the rotation (i.e the point where Emmy accelerates) is from the Earth, the more time the rotation will sweep through.

While it certainly does seem hard to accept that any acceleration, no matter how small, can cause such a large effect at sufficient distance, this is indeed what special relativity predicts. However, the idea becomes much more palatable once we acknowledge that the concept of simultaneity, while conceptually useful, isn't particularly physical. That is to say, the version of Albert Emmy considers simultaneous with herself jumping forwards by more than six years when she turns around isn't a problem, because Emmy's method of determining what is or isn't simultaneous with her isn't rooted in anything physically meaningful. If we want to take a more physical approach to analysing the twins paradox we should ask ourselves what each twin actually sees when they look at the other through a telescope.

We can answer this question using the relativistic Doppler effect. Let us start by considering what Emmy sees if she points her telescope towards the Albert. While she's flying away from the Earth, he will appear redshifted by a factor of $\sqrt{(5+4)/(5-4)} = 3$, meaning that he will appear to be ageing at one third the normal speed. Thus, Emmy will see Albert age by one year over the three years she spends on the outbound journey. Once she turns around, Emmy will now be heading towards Albert and so he will appear blueshifted by the same factor of three. As such, he will now appear to age at triple speed, meaning that Emmy will see him age a full nine years over her three year return journey. We now consider what Albert will see. Just as before, while Emmy is flying away from Earth she will appear redshifted by a factor of three. Furthermore, Emmy will appear redshifted to Albert until he sees her turn around, which occurs nine years after her departure. So Albert sees Emmy age three years over a nine year period. For the remaining year until her return, Emmy appears blueshifted and so Albert will watch her age a further three years. This is shown graphically in Fig. 10.

The nice thing about this approach is that it can predict the correct age difference from the point of view of each twin without the need for any infinities. It also gives us a nice way to think about how the acceleration affects the age difference between the two twins. In this approach the key asymmetry which leads Albert to be older than Emmy is that he has to wait until the light from her turning around reaches

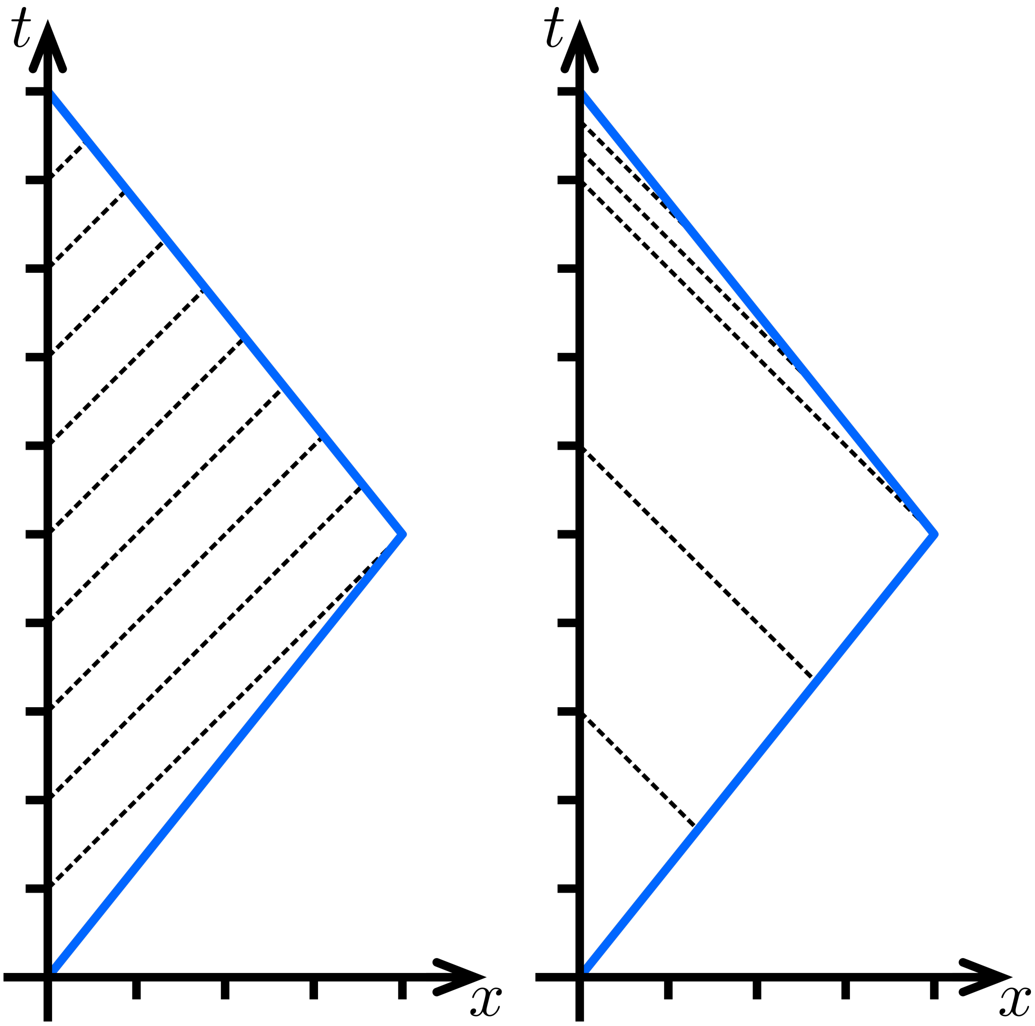


Figure 10. Spacetime diagrams of the twins paradox in the Earth's reference frame. On the left the dashed lines represent light rays emitted by Albert once every year. On the right the dashed lines are light rays emitted by Emmy once per year she experiences.

him before she transitions from red to blueshift, whereas Emmy sees Albert change as soon as she accelerates. This also aligns neatly with the fact that the size of the effect depends on Emmy's distance when she accelerates, because the farther away she is, the longer the time delay before Albert sees the acceleration.

There are two key points we should take away from our discussion of the twins paradox. Firstly, while velocity is relative, acceleration is absolute, and so we must always ensure that accelerated frames of reference are treated with appropriate care. Secondly, simultaneity is not a particularly physical concept for relativistic systems, and so we shouldn't worry too much if it appears to be doing strange things, so long as physical things like observations made using light rays remain sensible.

3.2 The barn-pole paradox

The set-up for the barn-pole paradox is as follows. Suppose an indestructible barn which is five metres long. We now take a ten metre long steel pole and fire it through the open door of the barn at a speed such that length contraction cuts its length in half. The moving pole will fit exactly inside the barn and so we should be able to to close the barn doors with it inside. The paradox arises when we consider the same situation from the pole's reference frame. In this frame, the pole will be ten metres long, while the barn is length contracted to only two and a half metres, so it seems impossible that the entire pole could fit inside the barn.

The resolution to this paradox lies in the lack of common simultaneity between the two reference frames. To begin with let us dispense with closing the barn doors and simply leave them open as shown in Fig. 11. In this case the pole will simply

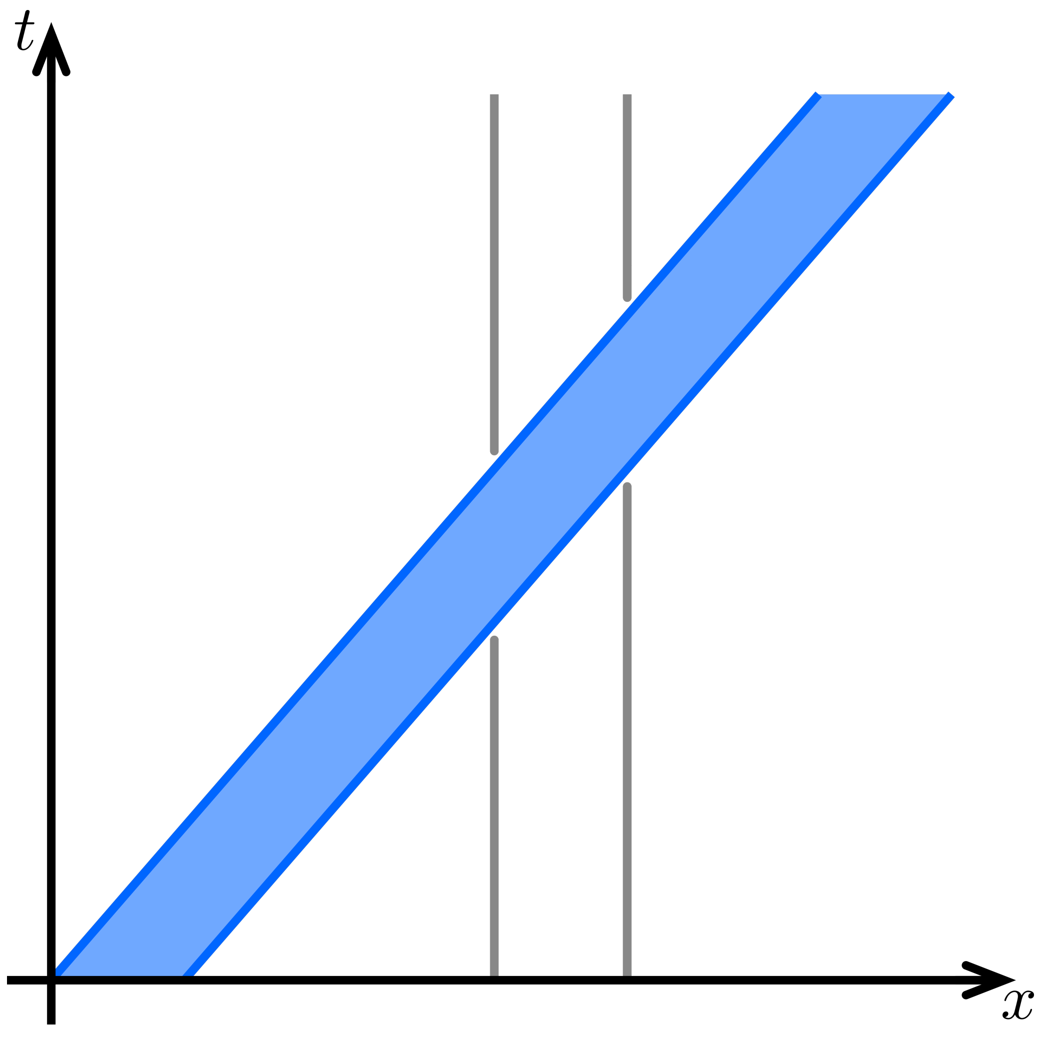


Figure 11. Spacetime diagram of the barn-pole paradox in the barn's frame. In the version depicted here the barn doors (grey) are open so the pole (blue) passes through unobstructed. Note that there is a brief moment of time where the pole is fully inside the barn.

pass through the barn. As it does so, there will be a single moment of time where the entire pole is simultaneously inside the barn. At this moment we could choose to quickly close and open the barn doors without affecting the pole in any way. If we now move into the pole's frame, these openings and closings will no longer be simultaneous. Instead, once the barn covers the first quarter of the pole the back door will close and then open, allowing the barn to continue moving over the pole until it covers the last quarter, at which point the front door will close and open.

Things become slightly more interesting if we now try to trap the pole in the barn by closing the doors while it is inside. Obviously, this is pretty unrealistic as a steel pole moving fast enough to have a Lorentz factor of two will have enough kinetic energy to tear through the doors no matter what they're made of. However, for the sake of our thought experiment we shall pretend that a material strong enough to

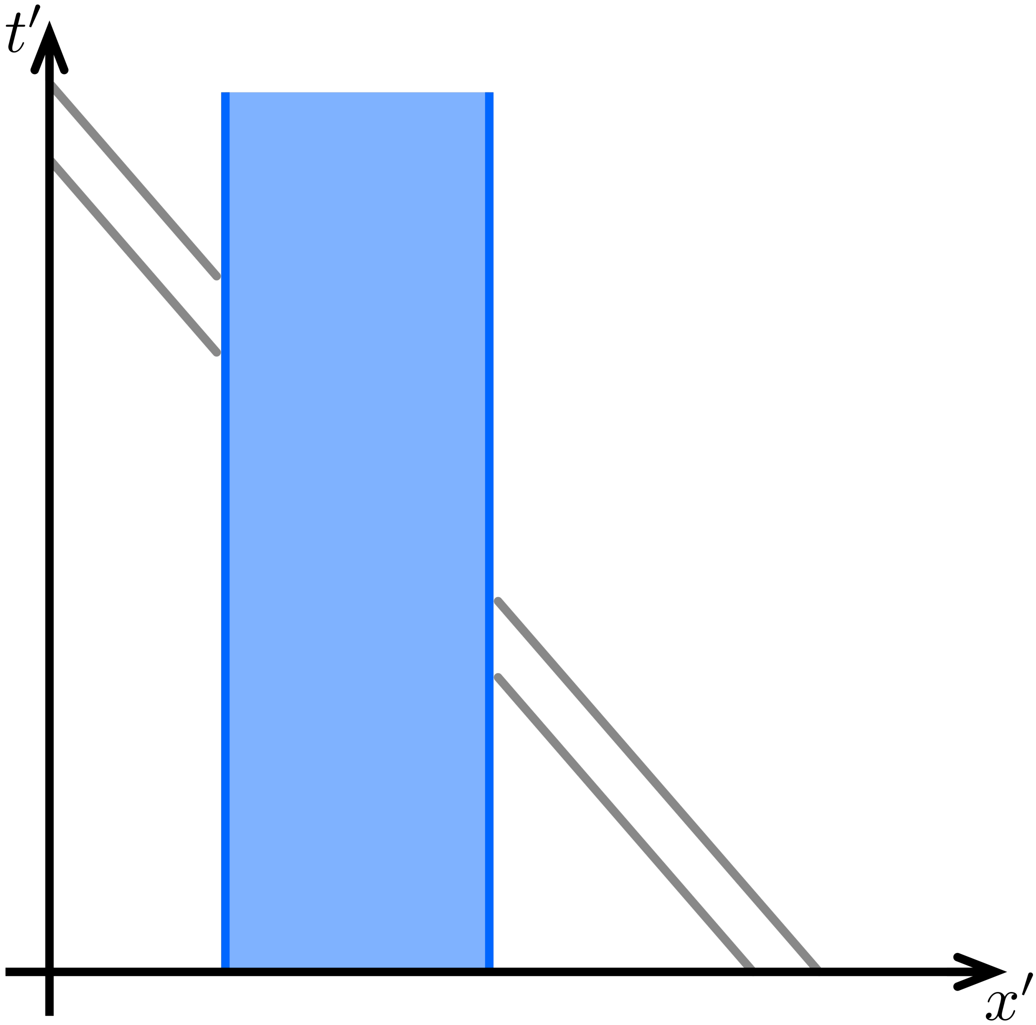


Figure 12. Spacetime diagram of the barn-pole paradox in the pole's frame. The end of the pole (blue) leaving the back door of the barn (grey) is not simultaneous with the other end of the pole entering the front door as it was in the barn frame. Thus, in this frame, the pole is never fully inside the barn at once, which makes sense given that the pole is four times longer than the barn.

withstand such an impact exists, and that the barn doors are made out of it. In the frame of the barn, the leading end of the pole will be brought to an immediate halt as soon as it slams into the closed back door. At that same moment the front door closes behind the trailing end of the pole as it enters the barn. However, the trailing end of the pole does not stop immediately; in order to respect causality it cannot be affected until light emitted by the initial collision reaches it. Thus, the pole will reach a minimum length even shorter than its five metre contracted length. Once it reaches this minimum length, the energy of the shock wave will slam into the end of the pole sending it flying back into the now closed door at nearly the speed of light. If we still insist that our barn is indestructible, then the kinetic energy will have nowhere to go except into the thermal energy of the highly compressed remains of the pole. This temperature will be more than sufficient to completely disintegrate the iron nuclei, leaving nothing but an ultra dense thermal soup of particles which is only prevented from exploding by the indestructible walls of the barn.

One nice way that we can understand the lack of common simultaneity between the two frames is as follows. Suppose that we are at rest relative to the barn, and have a single button that is wired into the barn doors. If we want both doors to close simultaneously, then we should position ourselves halfway between them such that the length of wire to the button is the same for each door. When we press the button, the electrical signal will propagate along the wires at light speed, reaching both doors and triggering them to close at the same time. However, things appear slightly differently in the frame of the pole. In this frame the signals from the button pressing still propagate out at the speed of light; however, the barn doors will now be in motion. The back door will be moving towards the signal, and so will receive it and close much earlier than the front door which is running away from its signal. Thus, we can see that the lack of simultaneity required to resolve the paradox follows directly from the invariance of the speed of light between reference frames.

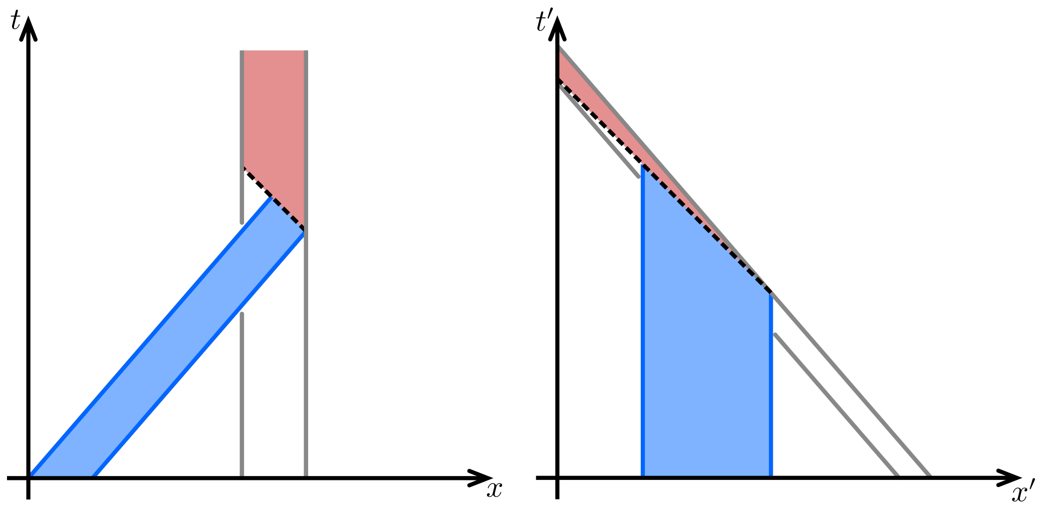


Figure 13. Spacetime diagrams of the barn-pole paradox with closed doors. The blue regions represent the pole in its natural unstressed state, and the red regions represent the contained explosion left behind after the pole collides with barn doors.

3.3 Bell's spaceship paradox

This paradox became famous after John Stewart Bell posed it as a question to physicists working at CERN during the seventies, and the majority of them got the answer wrong.[2] The paradox concerns the following thought experiment. Suppose two spaceships are initially floating at rest relative to one another, and separated by a distance of one kilometre. The two ships are connected by a kilometre long wire which has a tolerance of 1%, meaning it will snap if it is extended to more than 1% longer than its natural length. We now suppose that both spaceships simultaneously start accelerating, and maintain identical acceleration profiles until they reach a relativistic speed, say eighty percent the speed of light, when they cut their engines and continue floating. The question is: does the string joining the two ships break during the journey?

Our first instinct may be to argue that, since both ships have identical acceleration profiles, translational symmetry guarantees that the distance between the ships does not change, and so the string should not snap. However, after the acceleration has finished, the ships and string are moving at a substantial fraction of the speed of light, and so should be length contracted. That is to say, although the length of the string is still one kilometre in our original frame of reference, it must be significantly longer in its new rest frame, and so it should have snapped. The situation is much clearer when viewed in the eventual rest frame of the two ships, shown on the right in Fig. 14. In this frame, the ships are originally in motion until they decelerate

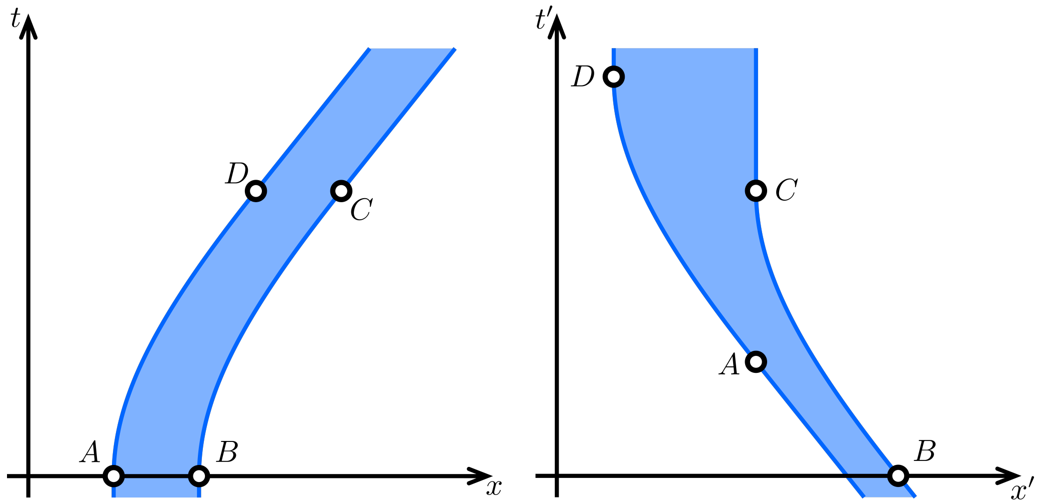


Figure 14. Spacetime diagrams of Bell's spaceship paradox. The scenario is shown in two inertial frames: on the left is $\mathcal{K}$, the initial rest frame of the ships, and on the right $\mathcal{K}'$, the final rest frame of the ships. The blue world lines represent the two ships, and the shaded region between them is the world sheet defined by the string joining the ships. The events A, B, C, D represent the beginning and end of the acceleration period for each ship.

[2] The CERN physicists all eventually arrived at the correct answer once they had taken a day or two to think the problem over, but for most of them, their first instinct was wrong.

to a stop. However, one of the ships starts decelerating later than the other, and during this delay it continues moving, stretching out the string between them until it eventually breaks.

The question of when exactly the string does snap is complicated, and will involve finding a relativistically valid equation describing the dynamics of the string, so we shall not concern ourselves with that here. The key point is: we know that the string cannot be unbroken and longer than 1.01 km when it is at rest, and by the principle of relativity this must apply in any inertial frame. Since we can see that the string would have to be longer than this once the two ships have come to rest in $\mathcal{K}'$, we can therefore conclude that it must have snapped at some point before that moment, even if we do not know exactly when.

3.4 Ehrenfest's paradox

Ehrenfest's paradox is based on the concept of a rotating rigid body. We imagine taking some disk of rigid material and spinning it up until points on the outer rim are moving at a substantial fraction of the speed of light. The argument then goes that, since the radius of the disk is always perpendicular to the velocity of the rotating disk, it should not be length contracted at all. On the other hand, the circumference of the disk is always parallel to the velocity, and so it certainly should be length contracted. The problem is that these two observations seem to imply that the ratio between the disk's circumference and radius is no longer equal to 2π, in opposition to elementary geometry.

The resolution of this paradox lies in the fact that special relativity does not permit perfectly rigid bodies to exist. In order to be consistent with causality, the speed of sound in a material must be subluminal, which places an upper limit on its stiffness. As such, it is perfectly acceptable for the circumference of the disk to be simultaneously length contracted and equal to 2π times the radius, this simply means that, because the circumference is longer than its natural length in its rest frame, it must be under stress. Of course, a rotating disk must already be under stress, since internal tension is required to prevent the disk from being ripped apart by centrifugal forces. For any physically reasonable material (i.e speed of sound much slower than light), these stresses are much larger than those generated by the length contraction of the circumference. The net effect of special relativity is to slightly reduce the expansion of the disk due to centrifugal forces, since some of the stress required to offset them can be generated via length contraction instead of increasing the radius of the disk.

This resolution does still leave the slight issue that, as measured by observers at rest on the disk, its circumference is greater than 2π times its radius. However, we can alleviate any concerns by noting that the frame of an observer at rest on the disk is non-inertial, since they must be accelerating to undergo circular motion, and so does not have to play by the rules of inertial frames; even the rules of Euclidean geometry. In fact, this realisation that non-inertial frames can display aspects of non-Euclidean geometry was instrumental in Einstein establishing his general theory of relativity.

4 Relativistic Dynamics

While studying the behaviour of high velocity rulers and clocks is interesting, it is not the sort of thing that usually comes to mind when thinking about physics. Instead, one usually pictures colliding billiard balls, masses on springs, and rockets. These systems are typically analysed using Newton's laws of motion; however, these laws are fundamentally inconsistent with the special theory of relativity.

For example, consider the scenario shown in Fig. 15. In an inertial frame $\mathcal{K}'$ two particles with equal mass collide head on at speed v, and are deflected by an angle of ninety degrees. If we calculate the Newtonian momentum before and after the collision we see that

$$p_x'^{(\text{before})} = mv - mv = 0\,,\; p_y'^{(\text{before})} = 0\,, \quad p_x'^{(\text{after})} = 0\,,\; p_y'^{(\text{after})} = mv - mv = 0\,, \quad (4.1)$$

and so momentum is conserved in this frame. We can use our knowledge of relativistic velocity addition to view this same collision in a new frame $\mathcal{K}$, where one of

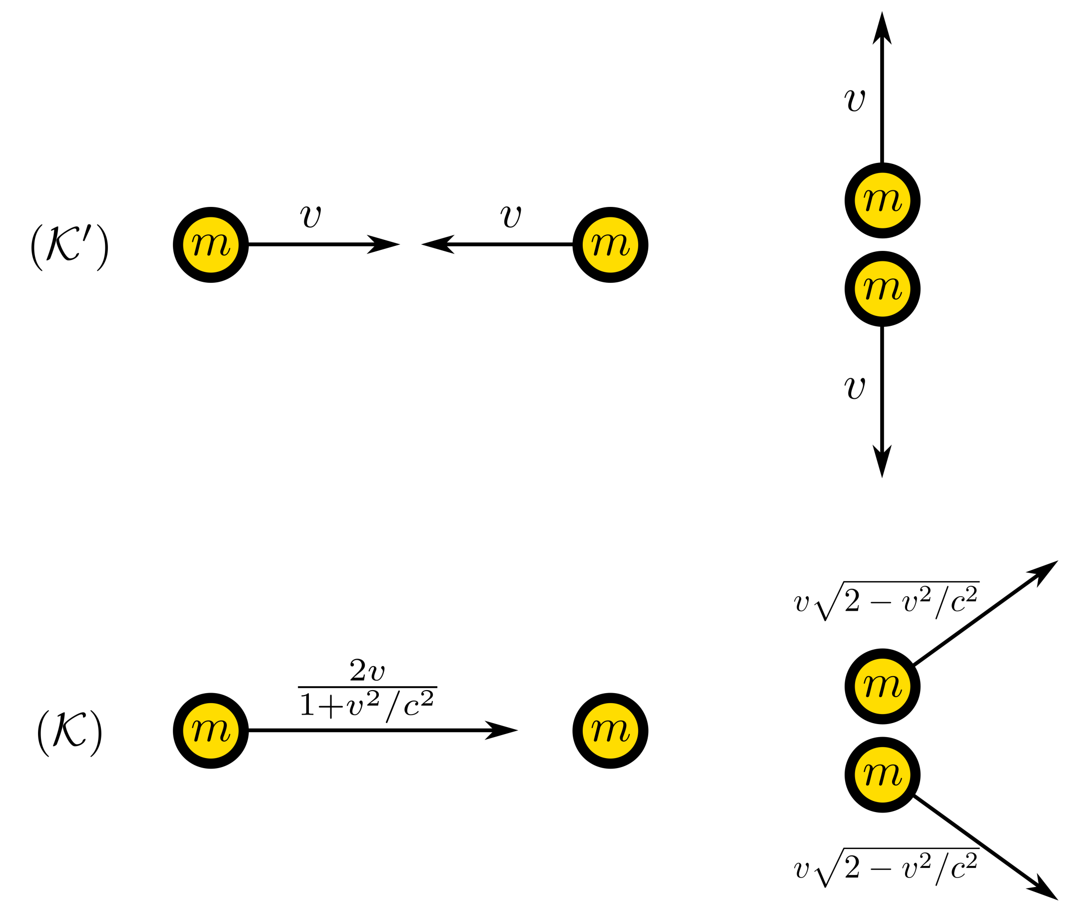


Figure 15. An elastic collision which does not conserve Newtonian momentum once we account for special relativity. One easy way to see that momentum is not conserved in $\mathcal{K}$ is to note that the angle between the two final velocities is slightly acute, while it is known that, in Newtonian mechanics, elastic collisions with a stationary target of equal mass always lead to perpendicular final velocities.

the particles is initially at rest. Carrying out the appropriate calculations, we find that the total momentum before the collision is

$$p_x^{(\text{before})} = \frac{2mv}{1+v^2/c^2}\,, \qquad p_y^{(\text{before})} = 0\,, \tag{4.2}$$

while the total momentum after is

$$p_x^{(\text{after})} = mv + mv = 2mv\,, \quad p_y^{(\text{after})} = mv\sqrt{1-v^2/c^2} - mv\sqrt{1-v^2/c^2} = 0\,. \tag{4.3}$$

The conclusion we are forced to accept is that Newtonian momentum conservation does not hold in all inertial reference frames, and thus, by the principle of relativity, it cannot be a true law of nature. Of course, abandoning Newtonian dynamics then raises the question of what new relativistic laws of motion we should embrace instead.

Technically speaking, special relativity cannot tell us what the correct laws of motion are; it can only tell us whether a particular set of laws are possible or not. If there are multiple different sets of laws which are consistent with the principle of relativity, then we must turn to experiment to determine which is correct. Fortunately, it turns out that relativistic consistency is a rather strict requirement, and so we can still deduce quite a lot about what the possible laws must look like. For example, we might try to salvage conservation of momentum by modifying the expression $p = mv$ into something more complicated. Special relativity alone is insufficient to tell us if this is possible, it may simply be that the universe does not possess any conserved quantity associated with translational motion; however, as we shall see, special relativity does tell us that, if such a quantity exists, there is only one possibility for what it could be.

4.1 Energy and momentum

The key to understanding momentum in special relativity is to understand how it transforms between frames. Let us start by considering a single particle which moves with velocity (u_x, u_y, u_z) in some inertial frame $\mathcal{K}$. Since a point particle has rotational symmetry about the axis of its motion, the particle's momentum must be parallel to its velocity. As such, we can say that

$$(p_x, p_y, p_z) = \big(\mathcal{M}u_x, \mathcal{M}u_y, \mathcal{M}u_z\big)\,. \tag{4.4}$$

The (potentially frame dependent) coefficient of proportionality $\mathcal{M}$ is sometimes referred to as the relativistic mass of the particle, since it is playing the role taken by the mass m in the Newtonian expression for momentum. We shall use this terminology for now, but note that it is usually considered somewhat old fashioned. We now consider a second frame $\mathcal{K}'$ in the standard configuration with $\mathcal{K}$. Using identical reasoning we can conclude that

$$(p'_x, p'_y, p'_z) = \big(\mathcal{M}'u'_x, \mathcal{M}'u'_y, \mathcal{M}'u'_z\big)\,. \tag{4.5}$$

We already know how the velocities in the two frames are related to one another, since they will simply obey the relativistic velocity addition formulae

$$(u_x, u_y, u_z) = \left(\frac{u'_x + v}{1+vu'_x/c^2}, \frac{u'_y}{\gamma_v(1+vu'_x/c^2)}, \frac{u'_z}{\gamma_v(1+vu'_x/c^2)}\right)\,. \tag{4.6}$$

We can use this fact to determine how the momentum transforms as follows. Firstly, we note that, since the momentum is parallel to the velocity, we must have[1]

$$\frac{p_x}{p_z} = \frac{u_x}{u_z} = \frac{\gamma_v(u'_x + v)}{u'_z} = \frac{\gamma_v(p'_x + v\mathcal{M}')}{p'_z}\,, \quad \frac{p_y}{p_z} = \frac{u_y}{u_z} = \frac{u'_y}{u'_z} = \frac{p'_y}{p'_z}\,. \tag{4.7}$$

This implies that the transformation law for the components of the momentum should take the form

$$p_x = K\gamma_v(p'_x + v\mathcal{M}')\,, \quad p_y = Kp'_y\,, \quad \text{and} \quad p_z = Kp'_z\,, \tag{4.8}$$

for some coefficient of proportionality K, which we will now determine. Since momentum is an additive quantity, i.e the total momentum of a composite system is the sum of the momenta of its subsystems, the un-primed momenta should be linear functions of their primed counterparts so that this additive property holds in both inertial frames. Thus, K cannot depend on the momentum, and must be solely determined by the relative velocity v between the frames. Furthermore, since all directions are equivalent, K can only depend on the magnitude of the velocity and not its direction. This implies that K is the same for the transformation from $\mathcal{K}$ to $\mathcal{K}'$ as it is from $\mathcal{K}'$ to $\mathcal{K}$. Carrying out both of these transformations brings us back to where we started and so $K^2 = 1 \implies K = 1$.[2] We can also deduce the transformation law for the relativistic mass using

$$\frac{\mathcal{M}}{\mathcal{M}'} = \frac{p_z u'_z}{p'_z u_z} = \gamma_v(1 + vu'_x/c^2)\,. \tag{4.9}$$

Putting this all together leads us to final result that

$$\mathcal{M} = \gamma_v\Big(\mathcal{M}' + vp'_x/c^2\Big)\,, \quad p_x = \gamma_v\Big(p'_x + v\mathcal{M}'\Big)\,, \quad p_y = p'_y\,, \quad \text{and} \quad p_z = p'_z\,. \tag{4.10}$$

Interestingly, this takes the exact same form as the Lorentz transformation for the spacetime coordinates. We already know that the components of the momentum will mix together under rotations in exactly the same way as the spatial coordinates, because the direction of the momentum vector comes directly from the velocity which is based on the spatial coordinates. Since any Lorentz transformation can be composed as a mixture of rotations and a boost in the standard configuration, this tells us that $\mathcal{M}, p_x, p_y, p_z$ behave identically to t, x, y, z under any Lorentz transformation. It follows then that the quantity

$$\mathcal{M}^2c^2 - \left(p_x^2 + p_y^2 + p_z^2\right) = m^2c^2\,, \tag{4.11}$$

must be frame invariant, for the same reasons as the invariance of the spacetime interval. We can see that the m on the right hand side must be the particle's mass by recognising that we should recover Newtonian mechanics in the limit of very small velocities, and so in the particle's rest frame, where its momentum is zero, $\mathcal{M}$ should just be equal to its mass.

[1] Technically speaking, we are assuming here that $p_z \neq 0$; however, we can always imagine making it some finite quantity and later taking the limit as it tends to zero, so this is not a huge oversight.
[2] We choose the positive root because physically the transformation should be a continuous function of of v.

Since $p_x^2 + p_y^2 + p_z^2 = \mathcal{M}^2 u^2$, where u is the speed of the particle, we can use (4.11) to find expressions for the relativistic mass and hence the momentum. This yields

$$\mathcal{M} = \gamma_u m\,, \quad (p_x, p_y, p_z) = (\gamma_u m u_x, \gamma_u m u_y, \gamma_u u_z)\,. \tag{4.12}$$

Although we derived momentum transformations for a single particle, because they are linear, they must apply equally well to the total momentum and relativistic mass summed over an entire system. Furthermore, since the momenta in $\mathcal{K}$ are linear combinations of their primed counterparts and the relativistic mass, the changes in the un-primed momenta will be linear combinations of the changes in momenta and relativistic mass in $\mathcal{K}'$. As such, for momentum to be conserved in $\mathcal{K}$ it is not sufficient for momentum to be conserved in $\mathcal{K}'$, we must also conserve relativistic mass. However, due to the linearity of (4.10), if we conserve momentum and relativistic mass in one frame, they will automatically be conserved in all frames. Thus, if we want to make momentum conservation a law of physics, we can, but only if we make conservation of relativistic mass a law as well. Since a particle's relativistic mass is approximately equal to its ordinary mass when it moves at speeds much slower than light, this explains why mass is conserved in Newtonian mechanics.

There is another conserved quantity that can depend on a particle's motion, and is as of yet unaccounted for: energy. In general, trying to make energy conservation frame invariant so that it becomes compatible with the principle of relativity will necessitate the introduction of further conservation laws in the same way as momentum conservation required the conservation of relativistic mass. This is problematic because we do not know of any other conserved quantities[3], and it seems unlikely that we would fail to notice any if they existed. The trick to avoiding this is to ensure that energy can be constructed from the conserved quantities we already have. The only way of doing this while maintaining the additivity of energy and its independence on the direction of a particle's motion is to define

$$E = \mathcal{M}c^2 = \gamma_u m c^2\,, \tag{4.13}$$

where the factor of c^2 is chosen to obtain the correct units. We shall investigate this relation in more detail later, but for now we shall simply accept it as true. This relationship is the reason why the concept of relativistic mass is considered outdated; nowadays it is more common to refer only to the energy and momentum. As such, it is worth quickly rewriting our previous results in terms of these variables. The transformation law between reference frames becomes

$$E = \gamma_v\Big(E' + v p_x'\Big)\,, \quad p_x = \gamma_v\Big(p_x' + vE'/c^2\Big)\,, \quad p_y = p_y'\,, \quad \text{and} \quad p_z = p_z'\,, \tag{4.14}$$

while the invariant expression for the particle's mass becomes

$$E^2 - \big(p_x^2 + p_y^2 + p_z^2\big)c^2 = m^2 c^4\,, \tag{4.15}$$

which is known as the relativistic mass-energy-momentum relation.

[3] There is also angular momentum, but this can be ruled out because angular momentum depends on the location of the spatial origin, whilst energy does not, and so there is no way the two could mix under changes of reference frame.

We remarked earlier that the transformation law for energy and momentum applies equally well to a whole system as it does for a single particle. As such, the expression

$$\left(\sum E\right)^2 - \left(\sum p_x\right)^2 c^2 - \left(\sum p_y\right)^2 c^2 - \left(\sum p_z\right)^2 c^2 = M^2 c^4 \,, \tag{4.16}$$

is frame invariant when evaluated for any particular system. We refer to M as the invariant mass of that system, and it can be interpreted as the energy carried by the system in a frame where it has zero momentum, divided by c^2. In general, the invariant mass will not be equal to the total mass of all the particles present, but it will obey the inequality

$$M \geq \sum_{\text{particles}} m_i \,, \tag{4.17}$$

with equality if and only if all the particles are at rest with respect to one another. We can see why this inequality must hold by once again evaluating (4.16) in a frame where the total momentum is zero. Since a particle's energy must always be greater than or equal to its mass times c^2, the total energy in the zero momentum frame must be greater than or equal to the sum of the masses times c^2, and thus the invariant mass is greater than or equal to the sum of the individual masses.

For a more explicit demonstration, let us consider a system of two particles, which have a relative velocity u_{rel}. Choosing to evaluate (4.16) in a frame where one of the particles is at rest, we find that

$$\left(m_1 c^2 + \gamma_{u_{\text{rel}}} m_2 c^2\right)^2 - \left(\gamma_{u_{\text{rel}}} m_2 u_{\text{rel}}\right)^2 c^2 = M_{1,2}^2 c^4 \,. \tag{4.18}$$

Collecting the terms together, and using the fact that $\gamma_u^2 (c^2 - u^2) = c^2$, we can rearrange this to give

$$m_1^2 c^4 + m_2^2 c^4 + 2 m_1 m_2 c^4 \gamma_{u_{\text{rel}}} = M_{1,2}^2 c^4 \,, \tag{4.19}$$

which in turn can be used to conclude that the invariant mass is given by

$$M_{1,2} = \sqrt{(m_1 + m_2)^2 + 2 m_1 m_2 \left(\gamma_{u_{\text{rel}}} - 1\right)} \geq m_1 + m_2 \,. \tag{4.20}$$

4.2 Massless particles

One interesting possibility that relativity opens up is the existence of particles with zero mass. In Newtonian mechanics, such a particle would not carry any momentum or kinetic energy, and so essentially does not exist. The reason for this is that massless particles can only move at the speed of light, and so are inherently incompatible with the Newtonian limit where all speeds are much slower than light. We can see this as follows. First we recall that momentum is related to velocity according to

$$(p_x, p_y, p_z) = \left(\frac{E u_x}{c^2}, \frac{E u_y}{c^2}, \frac{E u_z}{c^2}\right) , \tag{4.21}$$

which implies that a particles speed can be determined from its energy and momentum according to the relationship

$$u = \frac{p c^2}{E} \,. \tag{4.22}$$

It follows from the mass-energy-momentum relationship (4.15) that for a massless particle $E = pc$, and so its speed must be equal to the speed of light.

One interesting thing to note is that, for a massless particle travelling along the x axis, we can set $p_x = E/c$, which leads the energy transformation law to take the form

$$E = \gamma_v \Big(E' + vE'/c \Big) = \frac{c+v}{\sqrt{c^2 - v^2}} E' = \sqrt{\frac{c+v}{c-v}} E' \,, \tag{4.23}$$

which is an identical expression to the Doppler shift for frequencies. This makes sense as photons, which move only at the speed of light, would have to be massless particles, and we know that the energy of a photon is related to its frequency.

4.3 Why does $E = \mathcal{M}c^2$?

It is all well and good us identifying the relativistic mass with energy; however, we should probably take some time to justify that this association does indeed make sense. For example, let us take a Taylor series expansion of the energy for a particle of mass m and speed u

$$E = \gamma_u mc^2 = \frac{mc^2}{\sqrt{1 - v^2/c^2}} = mc^2 + \frac{mu^2}{2} + \mathcal{O}(u^4/c^4) \,. \tag{4.24}$$

This is a very encouraging result as we can see that, ignoring the constant rest energy of mc^2, this reproduces the Newtonian result for kinetic energies at speeds much slower than light. A further justification of this result can be found if we consider the work done to accelerate a particle from rest to speed u. Assuming for simplicity that the force accelerating the particle is always parallel to its velocity, the work done on the particle, and hence its kinetic energy should be given by

$$E_K = \int F \, dx \,, \tag{4.25}$$

where F is the force acting on the particle. If we take the force to be given by the time derivative of momentum, then we will find that

$$E_K = \int \frac{dp}{dt} \, dx = \int \frac{dp}{dt} \frac{dx}{dt} \, dt = \int \frac{dx}{dt} dp = \int u \, dp \,. \tag{4.26}$$

With a little algebra we can obtain an expression for the particle's velocity as a function of its momentum, allowing the final integral to be written as

$$E_K = \int \frac{pc}{\sqrt{p^2 + m^2c^2}} \, dp \,. \tag{4.27}$$

This integral is most easily evaluated by recognising that the integrand can be written as a derivative to obtain

$$E_K = \int \frac{d}{dp} \Big(\sqrt{p^2c^2 + m^2c^4} \Big) \, dp. \tag{4.28}$$

Evaluating this integral, and taking care with the lower limit which corresponds to the particle at rest and therefore $p = 0$, we finally obtain

$$E_K = \sqrt{p^2c^2 + m^2c^4} - mc^2 = (\gamma_u - 1)mc^2\,. \tag{4.29}$$

This is also in agreement with our identification of the total energy as $E = \gamma_u mc^2$, so long as we accept that an object carries an energy of mc^2 even when it is at rest.

There is however an interesting distinction to be drawn between the relativistic and Newtonian cases. Inelastic collisions are collisions where, on first glance, energy is seemingly not conserved. These are not just theoretical, they make up the vast majority of collisions observed in real life. The Newtonian explanation is that energy is still conserved, it is just dissipated into forms which we cannot observe as easily as macroscopic kinetic energy, such as the microscopic kinetic energy associated with thermal motion, or potential energy in chemical bonds. However, in relativistic mechanics the conservation of the energy given by $\gamma_u mc^2$ is essential for momentum conservation, and so it must include all possible forms the energy could possibly take. We have already established that the macroscopic kinetic energy is given by $(\gamma_u - 1)mc^2$, and so the remaining rest energy mc^2 must include all the thermal and potential energies.

We can convince ourselves of this fact through the following argument. Let us consider a of mass m which spontaneously decays into a pair of photons (this could happen, for example, to a neutral pion). In the rest frame of the particle $\mathcal{K}'$, momentum conservation requires that the two photons be equal and opposite. Energy conservation then requires that each must carry an energy $E_0/2$, where E_0 is the total rest energy of the particle. If we orient our axes such that the photons travel along the x' direction, and boost to a new frame $\mathcal{K}$, as shown in Fig. 16, the total momentum of the photons will be given by

$$p_x = \frac{E_0}{2c}\sqrt{\frac{c+v}{c-v}} - \frac{E_0}{2c}\sqrt{\frac{c-v}{c+v}} = \frac{E_0 v}{c\sqrt{c^2 - v^2}} = \frac{\gamma_v E_0 v}{c^2}\,. \tag{4.30}$$

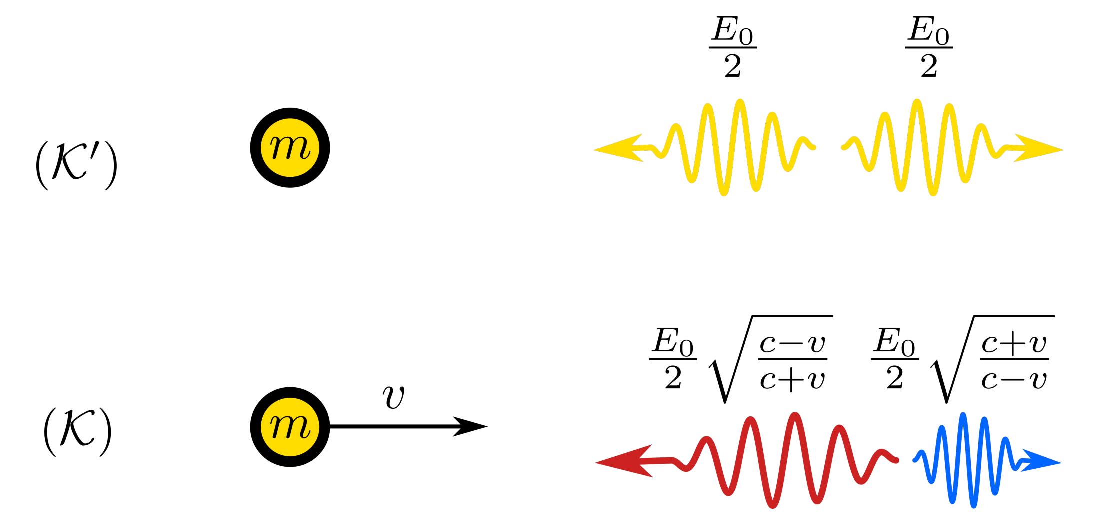


Figure 16. A massive particle decaying into a pair of photons.

Comparing this expression to the momentum of the massive particle, $p_x = \gamma_v mv$, it is clear that, if momentum is to be conserved in $\mathcal{K}$, the rest energy of the particle must be given by $E_0 = mc^2$.

This revelation is so significant that it is worth repeating. We have shown that, if the laws of energy and momentum conservation are to hold in the framework of special relativity, an object's inertial mass must reflect its energy content. In other words, if we were to take two otherwise identical objects, and transfer energy to one of them by heating it up, there would then be a small mass difference between them. This mass difference is very small; it would take nearly 10^{17} J of energy to produce a mass difference of only 1 kg. There are however cases where this effect becomes significant. For example, when an electron and positron annihilate to produce a pair of photons, all of the energy stored in their masses is converted into electromagnetic energy, which can eventually be thermalised by the surroundings.

4.4 Pion decay

We shall now demonstrate how to apply the relativistic expressions for energy and momentum to a few physical examples. The first shall be the kinematics involved in the decay of a charged pion into a muon and a neutrino, as shown in Fig. 17. In particular, our goal is to determine the energies of the muon and neutrino, in the pion's rest frame.

A good first step in approaching any calculation is to simply count the number of variables and number of constraints. The final state has a total of eight degrees of freedom, an energy and three components of momentum for each particle. However, we can immediately impose two constraints: the mass-energy-momentum relations for each particle, leaving six degrees of freedom. We know the energy and momentum of the initial state (since the pion is at rest it has energy $m_\pi c^2$ and zero momentum) and so energy and momentum conservation allow us to impose four further constraints. The remaining two degrees of freedom correspond to the two angles needed to specify the direction of one of the particles. Since rotations in space will not affect the energies of the two particles, it should thus be possible to determine these uniquely.

We shall let E and p denote the energy and magnitude of the muon's momentum respectively. The muon mass is known to be $m_\mu c^2 = 106\,\mathrm{MeV}$, and we know that it must be given by the mass-energy-momentum relation

$$E^2 - p^2c^2 = m_\mu^2 c^4\,. \tag{4.31}$$

For simplicity, we shall treat the neutrino as massless. Strictly speaking, this isn't true since neutrinos are known to have non-zero, albeit very small, masses. However, the neutrino mass is many orders of magnitude smaller than the pion mass, and so the overwhelming majority of the energy carried by the neutrino is kinetic rather than its rest energy. In this circumstance the neutrino mass is negligible, and so we can approximate it as a massless particle. Conservation of momentum implies that the magnitude of the neutrino's momentum must also be p, so the massless approximation implies that the neutrino's energy is pc.

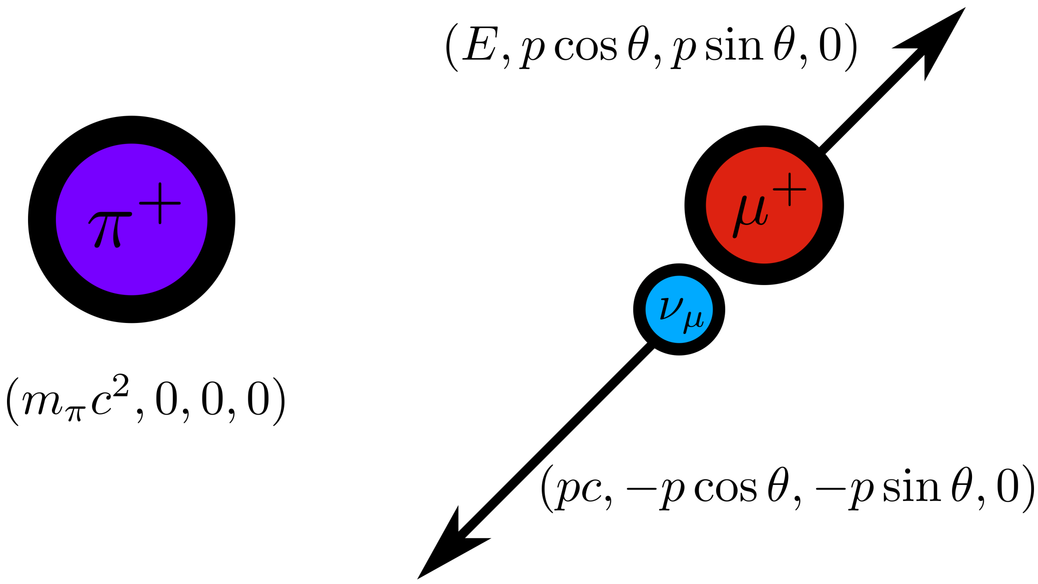


Figure 17. The dominant decay mode of the positive pion: $\pi^+ \to \mu^+\nu_\mu$, shown in its rest frame. Conservation of momentum ensures that the two daughter particles have momenta of equal magnitude p, which point in opposite directions. The neutrino is treated as massless so its energy is simply pc, while the muon has energy $E = \sqrt{p^2c^2 + m_\mu^2 c^4}$.

Since the initial energy is just the rest energy stored in the pion's mass ($m_\pi c^2 = 140\,\text{MeV}$), conservation of energy then tells us that we must have

$$E + pc = m_\pi c^2\,. \tag{4.32}$$

We now have two equations in the two unknowns E and p, so solving the problem is now just a matter of algebra. The most elegant way to proceed is to note that

$$E - pc = \frac{E^2 - p^2c^2}{E + pc} = \frac{m_\mu^2 c^2}{m_\pi}\,, \tag{4.33}$$

which now reduces the problem to two linear simultaneous equations. We now simply add (4.32) and (4.33) together to obtain

$$E = \frac{E + pc}{2} + \frac{E - pc}{2} = \frac{m_\pi c^2}{2} + \frac{m_\mu^2 c^2}{2m_\pi} = \frac{m_\pi^2 + m_\mu^2}{2m_\pi}c^2\,. \tag{4.34}$$

From here it is a relatively simple matter to obtain the corresponding expression for the momentum

$$pc = \frac{m_\pi^2 - m_\mu^2}{2m_\pi}c^2\,, \tag{4.35}$$

and the numerical answers

$$E = 110\,\text{MeV} \qquad pc = 30\,\text{MeV}\,. \tag{4.36}$$

4.5 Proton-electron capture

The weak interaction makes it possible to interconvert protons and neutrons. For example, the proton-electron capture reaction $p^+e^- \to n\nu_e$ converts a proton into a neutron whilst simultaneously converting an electron into a neutrino. However, protons come into close proximity with electrons all the time, and yet this reaction rarely occurs in everyday life, a fact which is pretty important for the continuation of said life. This is due to the mass difference between the final and initial particles (about $782\,\mathrm{keV}/c^2$), which needs to be compensated for by the kinetic energy of the electron and proton.

We want to determine under what conditions a collision between a proton and electron could possibly lead to proton-electron capture. Just as before, we can start by tallying up the available degrees of freedom. The initial state has eight parameters for us to specify (an energy and three momentum components for each particle); however, the mass-energy-momentum relations for the electron and proton reduce this to six. Furthermore, whether a particular process is possible or not is a law of nature, and so must be frame independent. Thus, we can always Lorentz transform into a particularly convenient frame, such as the proton's rest frame, removing three degrees of freedom. Finally, the possibility of a process should be unchanged by rotations, so we can choose to align our x axis with the electron's velocity, removing two more degrees of freedom. As such, the possibility of proton-electron capture can depend on only one parameter, which we can take to be the electron kinetic energy in the proton's rest frame.

Let us consider a collision between a stationary proton, and an electron with kinetic energy E_K, as shown in Fig. 18. We can calculate the minimum kinetic energy for proton-electron capture to take place as follows. First we note that, since the invariant mass is a function of the total energy and momentum, which are conserved, it too must be conserved in the interaction. We can calculate the invariant mass for

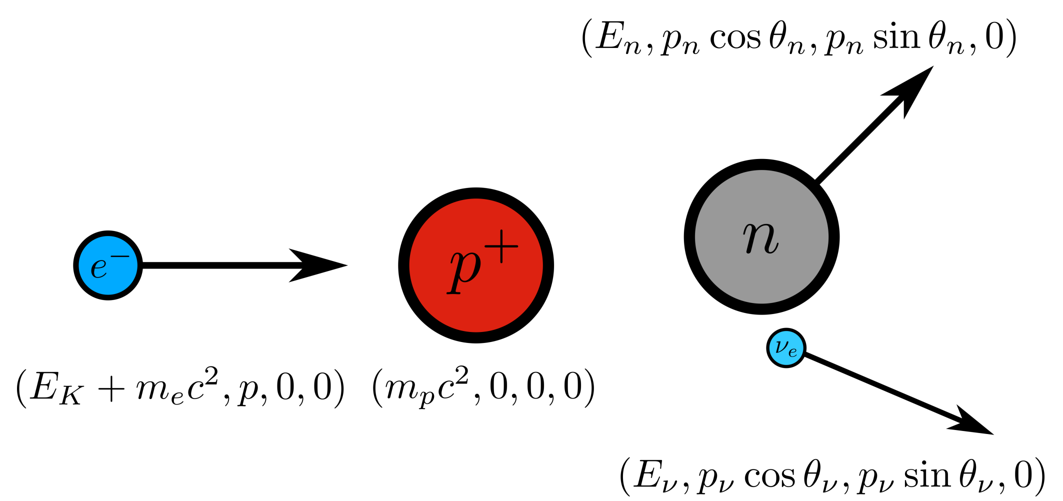


Figure 18. A high energy electron colliding with a stationary proton to produce a neutron and an electron neutrino.

the initial state to be

$$M^2 = (E_K/c^2 + m_e + m_p)^2 - p^2/c^2 = (m_e + m_p)^2 + 2m_p E_K/c^2 \,,$$

where p is the magnitude of the electron's momentum, which can be calculated by applying the mass-energy-momentum relation for the electron

$$p^2 = (E_K/c + m_e c)^2 - m_e^2 c^2 = E_K^2/c^2 + 2m_e E_K \,.$$

We know that the invariant mass of the final state must be greater than or equal to the sum of the individual particle masses, i.e

$$M^2 \geq (m_n + m_\nu)^2 \,,$$

and so in order for the reaction to occur we must have

$$E_K \geq \frac{(m_n + m_\nu)^2 - (m_p + m_e)^2}{2m_p} c^2 = 783\,\mathrm{keV} \,.$$

It is worth noting that this threshold kinetic energy is slightly larger than the rest energy difference between the final and initial particles. This is because the neutron and neutrino cannot be produced at rest: they will need some kinetic energy of their own to satisfy conservation of momentum.

4.6 Compton scattering

Compton scattering refers to a high energy photon being scattered by a stationary electron. Unlike Thomson scattering, which occurs when the energy of the scattered photon is much smaller than the electron's rest energy, Compton scattering changes the photon's wavelength as well as its direction of travel. Our task is to determine the change in wavelength as a function of the photon's scattering angle.

Once again, we can start by counting the number of variables and constraints available. Assuming we know the initial state exactly, there are eight degrees of freedom

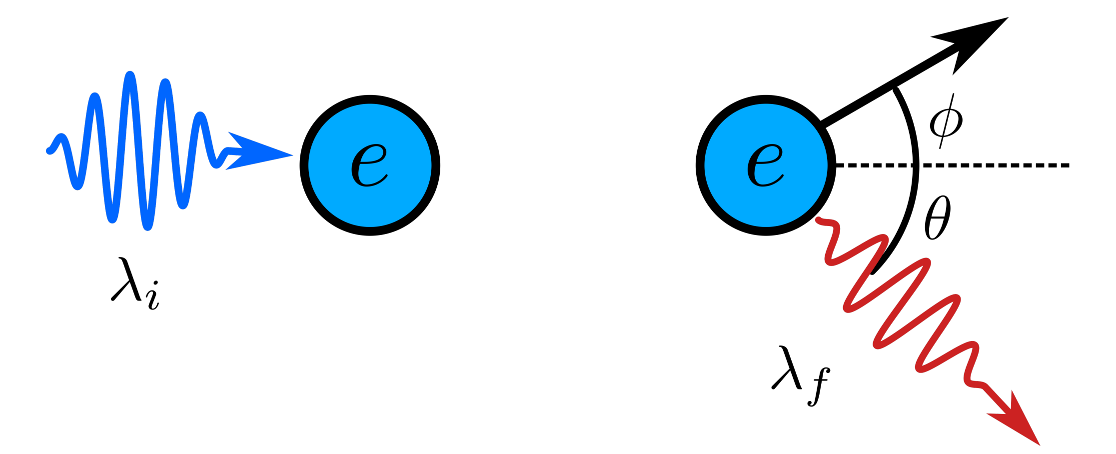


Figure 19. Compton scattering of a photon by a stationary electron.

in the final state; however, just as before this is reduced to six by the mass-energy-momentum relations for the electron and photon. Furthermore, conservation of energy and momentum impose four constraints to bring us down to two degrees of freedom. Unlike the pion decay, the initial state is not spherically symmetric, instead it only has rotational symmetry about the axis of the incident photon. Thus, we cannot determine the energy of the scattered photon uniquely; we can only express it as a function of the scattering angle θ.

We can now begin our more detailed analysis of the situation. We shall let E_i and E_f denote the energies of the incident and scattered photons respectively. Additionally, we shall let p denote the magnitude of the scattered electrons momentum. We are only interested in the photons and so a good guiding principle is to try and isolate p in any expressions so that it can be easily eliminated. For example, when we apply conservation of energy

$$E_i + m_e c^2 = \sqrt{p^2c^2 + m_e^2c^4} + E_f \implies p^2c^2 = (E_i + m_ec^2 - E_f)^2 - m_e^2c^4\,. \quad (4.37)$$

If we let the scattering angles of the photon and electron be θ and ϕ respectively, then conservation of momentum, together with the result $E = pc$ for photons, yields

$$pc\sin\phi = E_f \sin\theta \qquad E_i = E_f\cos\theta + pc\cos\phi\,. \quad (4.38)$$

We can eliminate ϕ by isolating the p terms, squaring them, and adding them together

$$p^2c^2 = p^2c^2\sin^2\phi + p^2c^2\cos^2\phi = E_f^2\sin^2\theta + (E_i - E_f\cos\theta)^2\,. \quad (4.39)$$

If we now equate the two different expressions for p^2c^2 in (4.37) and (4.39), the terms quadratic in E_f cancel and we are left with a linear equation which can easily be solved to give E_f as a function of E_i and θ

$$E_f = \frac{E_i m_e c^2}{m_ec^2 + E_i(1-\cos\theta)}\,. \quad (4.40)$$

We can now apply Planck's law that the energy of a photon is given by $E = hc/\lambda$ to convert this expression into one relating the wavelengths of the incident and scattered photons. This yields

$$\lambda_f = \lambda_i + \lambda_e(1-\cos\theta)\,, \quad (4.41)$$

where λ_e is known as the Compton wavelength of the electron and is given by the expression

$$\lambda_e = \frac{h}{m_e c} \approx 2.43\,\mathrm{pm}\,. \quad (4.42)$$

While we have considered only stationary electrons, similar increases in wavelength will be observed when high energy photons scatter of slow moving electrons. This phenomenon is particularly important in the core of the Sun where high energy gamma photons are produced via nuclear fusion reactions. Compton scattering of these photons transfers their energy to electrons which can then collide with other electrons and atomic nuclei, eventually thermalising the energy. This serves two purposes. Firstly, the heating of the solar core is important in maintaining sufficient pressure for the Sun to avoid collapsing under its own gravity. Secondly, the repeated lengthening of the photon wavelengths prevents gamma photons from making it to the surface of the Sun where they could freely reign down on Earth.

4.7 Relativistic rockets

Our final example shall be an examination of how rockets behave in special relativity. Traditionally, a rocket generates thrust by rapidly combusting fuel in an enclosed environment so that the exhaust gases are expelled at a high velocity away from the main body of the rocket. Since momentum must be conserved, the rocket must accelerate as it does this to offset the momentum carried by the exhaust. There are other slightly more unconventional systems which can lay claim to the status of rocket. For example, a laser pointer functions almost exactly like a rocket, only instead of converting fuel into exhaust, it converts the chemical potential stored in its battery into photons, which are then ejected. Of course, a laser pointer is not a very strong rocket: it would take gigawatts of power to achieve Newtons of thrust. A slightly more practical, if futuristic idea is an antimatter rocket. This rocket would hypothetical work by mixing matter and antimatter together in its reaction chamber, and then focusing the photons produced by the resultant annihilation so that they all leave through the back of the rocket. What both have in common is that the 'exhaust' is composed of photons, and so they are unlikely to be correctly described by Newtonian physics.

For simplicity, we shall assume that the rocket travels in a straight line, so its acceleration is always parallel to its velocity. We start by considering the rocket in its instantaneous rest frame. We then consider an infinitesimal interval of time dt over which the rocket loses a mass $-dm$ (dm is taken to be negative such that $-dm$ is positive). Over this period, the rapidity of the rocket increases from zero to $d\xi$. We choose to work with the rapidity because, as we saw earlier, changes in rapidity are frame invariant. Still, it will be a worthwhile exercise to see exactly how this comes about explicitly. If we let u be the rocket's velocity in some inertial frame, and du_0 the infinitesimal change in velocity, as measured in the rocket's instantaneous rest frame, then relativistic velocity addition tells us that

$$du = \frac{u + du_0}{1 + u du_0/c^2} - u \,. \tag{4.43}$$

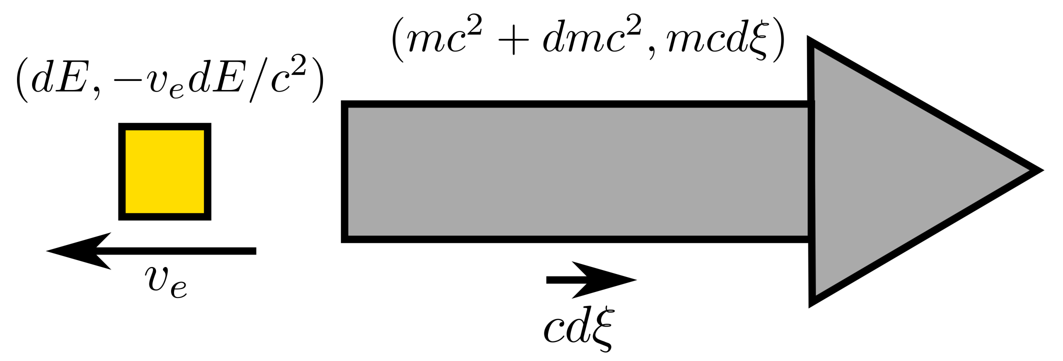


Figure 20. A rocket that has just ejected a small amount $-dm$ of its mass, shown in a frame where it was initially at rest. The exhaust is ejected from the rocket with a relative velocity of v_e.

Expanding (4.43) to first order in infinitesimal quantities we obtain

$$du = (1 - u^2/c^2)du_0 \implies du_0 = \frac{du}{1 - u^2/c^2} = d\left(\frac{c}{2}\ln\left(\frac{c+u}{c-u}\right)\right) . \tag{4.44}$$

Recalling that, by definition, the rapidity of the rocket ξ is such that $u = c\tanh\xi$, we can see that we must have

$$\xi = \operatorname{artanh}(u/c) = \frac{1}{2}\ln\left(\frac{c+u}{c-u}\right) , \tag{4.45}$$

and so $du_0 = cd\xi$. This relationship holds no matter what frame $d\xi$ is calculated in, and so $d\xi$ must be frame invariant as we desire. We can now determine how $d\xi$ is related to dm by simultaneously applying conservation of energy and momentum. Let the 'exhaust' ejected by the rocket carry an energy dE, as measured in the rocket's rest frame. Since the rocket's kinetic energy will be quadratic in its infinitesimal velocity it can be neglected, and so the energy carried by the rocket after losing the mass is simply $(m+dm)c^2$. The total energy before the ejection was simply the rest energy of the rocket, so energy conservation requires that

$$mc^2 = dE + (m + dm)c^2 \implies dE = -dmc^2 . \tag{4.46}$$

To first order in infinitesimal quantities, the momentum carried by the rocket is simply $mdu_0 = mcd\xi$. On the other hand, the momentum carried by the 'exhaust' will be given by $v_e dE/c^2$ in the opposite direction, where v_e is the relative speed with which the 'exhaust' is ejected. If, as in a realistic rocket, the 'exhaust' is not all ejected with a uniform velocity, then some appropriate average would have to be taken. For our purposes we shall simply assume that v_e is known. Since the total momentum in rocket's instantaneous rest frame is zero, momentum conservation will then tell us that

$$mcd\xi = \frac{v_e dE}{c^2} = -v_e dm \tag{4.47}$$

From here, we divide by m and then integrate both sides of (4.47) in order to end up at the final result we desire

$$\xi = \frac{1}{2}\ln\left(\frac{c+u}{c-u}\right) = \frac{v_e}{c}\ln\left(\frac{m_0}{m}\right) , \tag{4.48}$$

where m_0 is the mass the rocket had when at rest in the frame being considered. For comparison, the Tsiolkovsky rocket equation derived from Newtonian mechanics is

$$u = v_e \ln\left(\frac{m_0}{m}\right) . \tag{4.49}$$

As we might expect, this arises from (4.48) in the limit that $u \ll c$; however, interestingly, we do not need to make any assumptions about the size of v_e to recover the Newtonian result. This is not because Newtonian physics actually works for describing situations where v_e is comparable to c, it just so happens that two of the approximations that are built into the Newtonian calculation, namely the conservation of mass and the non-relativistic expression for the exhaust momentum, end up cancelling out to provide a deceptively non-approximate answer.

5 Experimental Evidence for Special Relativity

The experimental evidence for special relativity available to us in the twenty first century is, quite frankly, overwhelming. Modern particle accelerators are capable of reaching energies on the order of tera-electronvolts, which are more sufficient to push most subatomic particles up to extremely relativistic speeds. Every time these accelerators are turned on, which is quite often, they test special relativity, and they have yet to find it in error. However, physicists in the early twentieth century were not nearly so fortunate as us, and had to try quite a bit harder to convince themselves that relativity really was true. In what follows, we shall review a selection of historical experiments, which were among the first to provide experimental verification for some of the key predictions of special relativity.

5.1 Aberration of starlight

The aberration of starlight is a phenomenon where stars appear to periodically change their position in the sky over the course of a year, first observed at the end of the seventeenth century. Aberration is distinct from parallax, which occurs due to the Earth's different positions in space throughout the year, as evidenced by the fact that the amount of aberration is independent of a star's distance from the Earth. The first explanation of the phenomenon was put forward in 1727 by James Bradley using a corpuscular theory of light. Once it became apparent that light was wave-like and not corpuscular, Thomas Young adapted this explanation by assuming that light propagated through a stationary aether.

For simplicity, let us consider the aberration of a distant star located directly above the Earth's orbital plane. According to Young, the luminiferous aether is stationary relative to the Sun, and so the light from the star should rain vertically down on the Earth. However, the Earth is moving through the aether, and so a telescope on Earth would have to be at an angle to catch the light, as shown in Fig. 21. Thus, the star's position will appear shifted by a slight angle

$$\tan\theta = \frac{v}{c}\,. \tag{5.1}$$

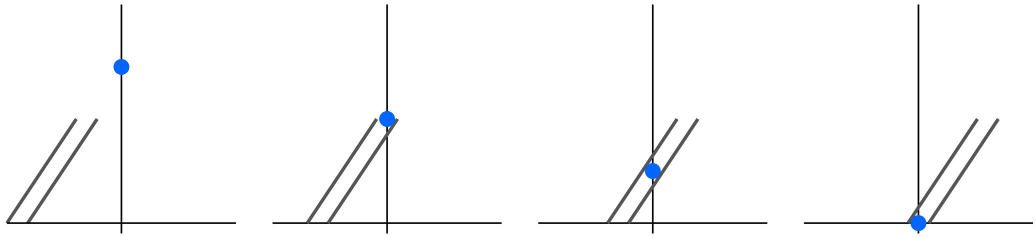


Figure 21. A moving telescope needs to be angled into the direction of its velocity to catch light which is moving straight down. Since the Earth's velocity changes periodically over the course of one year, this effect would cause the apparent angular position of stars to oscillate on a yearly cycle.

This result is consistent with observation, although it is worth bearing in mind that, for the Earth's orbit around the Sun, $v/c \approx 10^{-4}$, and so higher order corrections to the formula would not necessarily be measurable. The major problem with this explanation is that it cannot explain what happens when the telescope is filled with a refractive medium such as water. If we follow the same line of argument as before, then since the light will now fall more slowly through the telescope, it will need to be tilted at a steeper angle for the light to actually reach the bottom of the telescope. Unfortunately, experiments such as those by François Arago in 1810, showed that filling the telescope with water made no impact on the aberration angle.

By contrast, special relativity has no problem explaining both the aberration and its independence from the refractive index of the telescope. We start with an inertial frame $\mathcal{K}$ in which the Sun is at rest. We shall orient our axes such that the Earth's orbit is in the x, y plane, and the Earth's velocity points along the x axis. In this frame, the velocity of light from the star is simply c in the negative z direction. We can then apply the relativistic velocity transformation law to convert this into the Earth's rest frame $\mathcal{K}'$.

$$(u_x, u_y, u_z) = (0, 0, -c) \implies (u'_x, u'_y, u'_z) = (-v, 0, -c/\gamma_v)\,. \tag{5.2}$$

Thus, we can see that, to an observer in $\mathcal{K}'$, the incoming light makes an angle

$$\tan\theta = \frac{u'_x}{u'_z} = \frac{\gamma_v v}{c} = \frac{v}{c} + \frac{v^3}{2c^3} + \mathcal{O}(v^5/c^5)\,, \tag{5.3}$$

to the z axis. For the Earth's orbit, the relativistic correction to Bradley's formula will be roughly 5×10^{-13} rad. This is over an order of magnitude smaller than the precision possible with modern measurements such as the Gaia mission, so unfortunately we cannot directly rule out Bradley's formula.[1] However, the huge benefit of the relativistic approach is that, since the change in angle has nothing to do with the observation device, it will clearly be unchanged if we fill our telescope with water.

5.2 The Fizeau experiment

Given the length of time between the discovery of problems in Bradley and Young's explanation for stellar aberration and Einstein's development of special relativity, it is perhaps unsurprising that alternative explanations were available. In particular, Augustin-Jean Fresnel showed that if the luminiferous aether was partially dragged along by matter according to the relation

$$v_{\text{aether}} = \left(1 - \frac{1}{n^2}\right) v_{\text{matter}}\,, \tag{5.4}$$

where n is the refractive index of the matter, then he could successfully explain aberration in water filled telescopes. There are a number of potential problems with this seemingly simple formula. For example, the refractive index of a medium depends on the wavelength of light, so it is unclear at what velocity the aether should

[1] It is also worth noting that other effects in the solar system such as gravitational deflection of light will also produce corrections that are significant at that precision, and so would have to be accounted for.

actually move, or if it somehow moves at different speeds for each wavelength. A similar problem arises with birefringent materials, which have polarisation dependent refractive indices.

An alternative was a model put forward by George Stokes in which the aether would be completely dragged along by matter, and as such would obey $v_{\text{aether}} = v_{\text{matter}}$. Stokes' model explained aberration by assuming that, while the aether near the Earth was dragged along with it, the aether far away was at rest relative to the Sun. By making certain specific assumptions about the way the aether velocity varied over space, he was able to explain the deflection of starlight as a consequence of light passing through the velocity gradient in the aether. Since this had already occurred by the time the light reached the Earth, it would not depend on the telescope used to observe it. Stokes' model was not without problems. The most significant being that the model was not very specific as it he was not able to pin down the exact properties of the aether.

In 1851, Hippolyte Fizeau attempted to settle the debate between the two models by experimentally measuring aether drag. Fizeau did this by measuring the interference between two beams of light that were sent in opposite directions through flowing water, as shown in Fig. 22. He set up his apparatus so that the paths of the two beams would be of equal length. However, since the beams pass through the water in opposite directions, they would receive different contributions to their optical path length from any aether dragging that occurred. This difference in optical path length would lead to a phase difference between the two beams, which could be observed as a shift in the interference pattern formed when the two beams are combined.

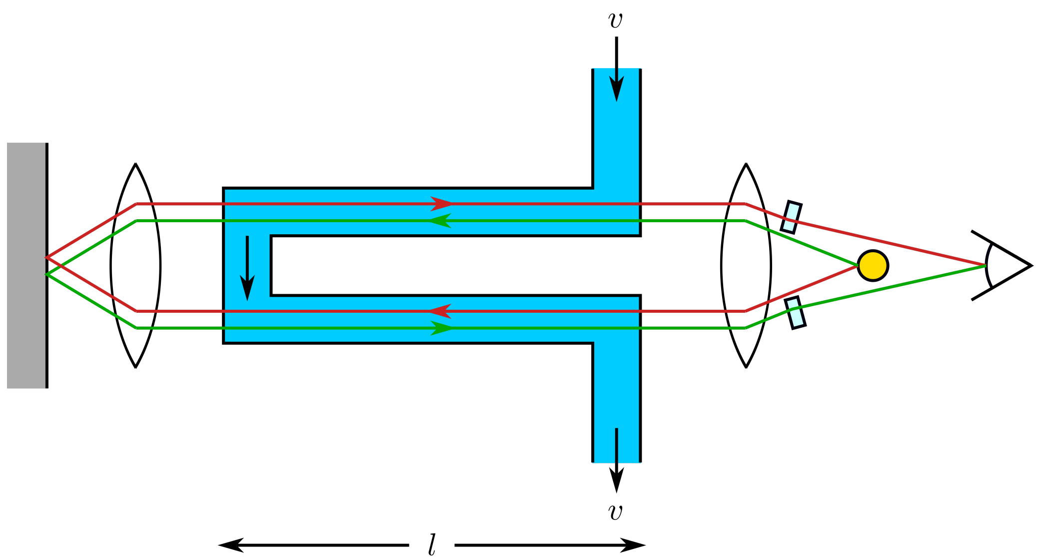


Figure 22. The schematic diagram of Fizeau's experiment. The two ray paths are shown in red and green, although in practice they will both be light of the same colour. Both paths are the same length, so the only phase difference that can develop is due to different speeds as the move through the flowing water. This phase difference will be detected as a translation of the observed diffraction fringes from their ordinary positions.

Let w_+ and w_- represent the speed of the rays as they move through the water with and against the flow respectively. The phase accumulated by a ray when it passes through one length of water is

$$\phi_\pm = \frac{lc}{\lambda w_\pm}\,, \tag{5.5}$$

where l is the length of the water, and λ is the free space wavelength of the light in question. Since each ray goes through the water twice, the total phase difference between the two paths will be

$$\Delta\phi = \frac{2l}{\lambda}\left(\frac{c}{w_-} - \frac{c}{w_+}\right)\,. \tag{5.6}$$

The values of $w_\pm$ will depend on the particular model in question. Applying Fresnel's aether drag formula (5.4), we obtain the result

$$w_\pm = \frac{c}{n} \pm \left(1 - \frac{1}{n^2}\right) v \implies \frac{c}{w_-} - \frac{c}{w_+} = \frac{2(n^2-1)v}{c} + \mathcal{O}(v^3/c^3)\,, \tag{5.7}$$

while using Stokes' complete dragging model gives

$$w_\pm = \frac{c}{n} \pm v \implies \frac{c}{w_-} - \frac{c}{w_+} = \frac{2n^2 v}{c} + \mathcal{O}(v^3/c^3)\,. \tag{5.8}$$

For any realistically obtainable velocity of water, the cubic and higher order terms will be negligible, and so the phase difference for each model can be expressed as

$$\Delta\phi_{\text{Fresnel}} = \frac{4(n^2-1)lv}{\lambda c} \quad \text{and} \quad \Delta\phi_{\text{Stokes}} = \frac{4n^2 lv}{\lambda c}\,. \tag{5.9}$$

For water, which has $n \approx 1.4$, the difference between the two models is roughly a factor of two so, as long as the experimental error can be made small compared to the size of the phase difference, it should be possible to distinguish the two models. Furthermore, since the wavelength of visible light is so short at $\sim 500\,\mathrm{nm}$, it should be possible to produce a significant phase difference with reasonably sized apparatus (on the order of $1\,\mathrm{m}$), and achievable water speeds (on the order of $10\,\mathrm{m\,s^{-1}}$). Much to Fizeau's surprise, when he actually carried out this experiment, his data fit far better with Fresnel's model than Stokes', providing experimental evidence that the aether is only weakly dragged by matter.

The special relativistic explanation of this experiment is as follows. In the rest frame of the water $\mathcal{K}'$, the light will propagate with speed c/n, since that is just the definition of the refractive index. We then simply transform this velocity into a frame $\mathcal{K}$ where the water has velocity v to deduce the speed of light through flowing water. Performing this calculation gives us

$$w_\pm = \frac{c/n \pm v}{1 \pm v/nc} = \frac{c}{n} \pm \left(1 - \frac{1}{n^2}\right) v + \mathcal{O}(v^2/c)\,. \tag{5.10}$$

For water speeds much slower than light, such as those used by Fizeau, this result is approximately that predicted by Fresnel, and as such is perfectly consistent with Fizeau's results.

5.3 The Michelson–Morley experiment

With the experimental verification of Fresnel's partial aether dragging model, a search began to try and measure the Earth's motion through the aether directly, without using external light sources like stars. The most famous such experiment was carried out by Albert Michaelson and Edward Morley in 1887. As shown in Fig. 23, they split a light beam between two perpendicular paths of equal length, then recombined the two beams and examined the resulting interference pattern. The two paths were set up so that one pointed along the direction of the Earth's velocity, and the other was perpendicular to it. The idea was that since the Earth, and hence the apparatus, was moving, the actual distance travelled through the aether would be different between the two paths, leading to a phase difference, and hence a shift in the interference pattern.

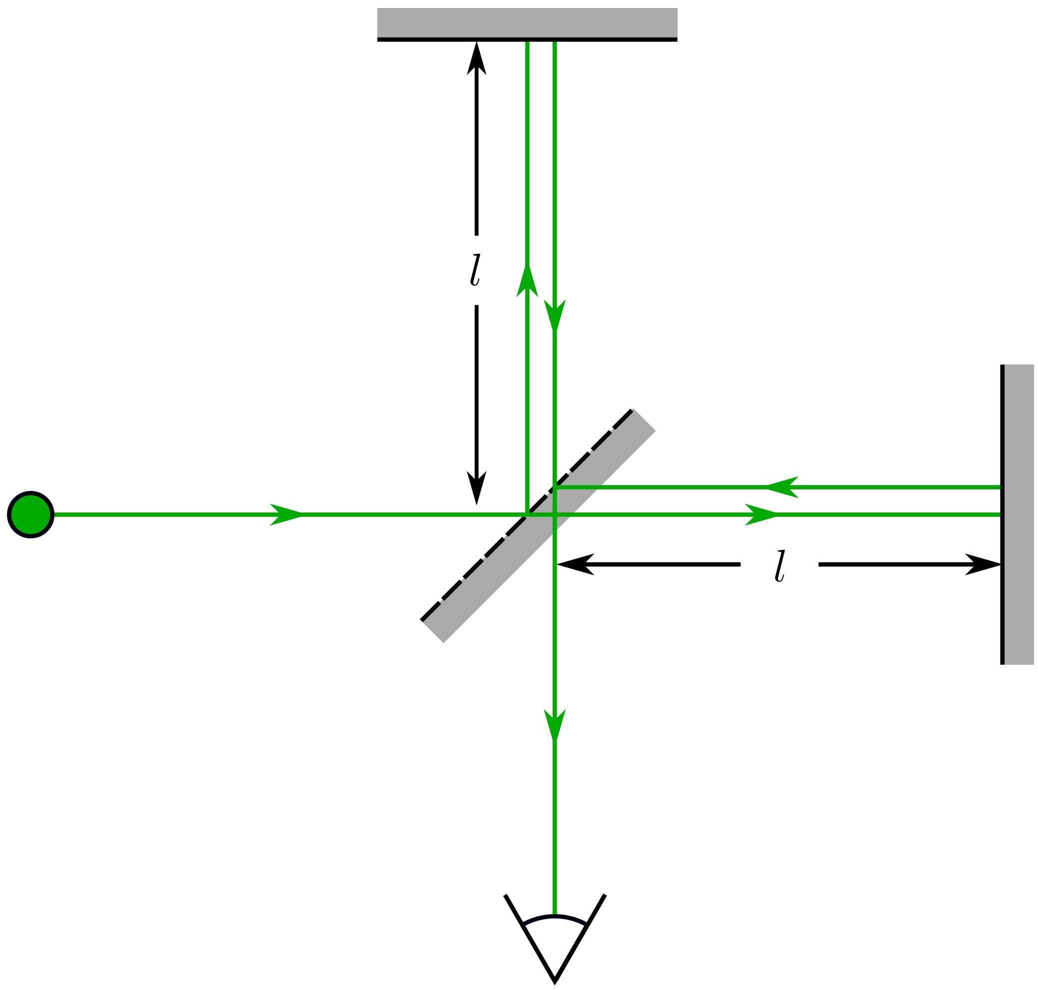


Figure 23. A simplified version of the apparatus used by Michelson and Morley for their aether drift experiment. A beam of light is sent into the apparatus from some source and hits the beam splitter in the centre. The two orthogonal beams then continue on until they are reflected back by the mirrors. The beams are then recombined by the beam splitter and sent towards the detector, where an interference pattern will form.

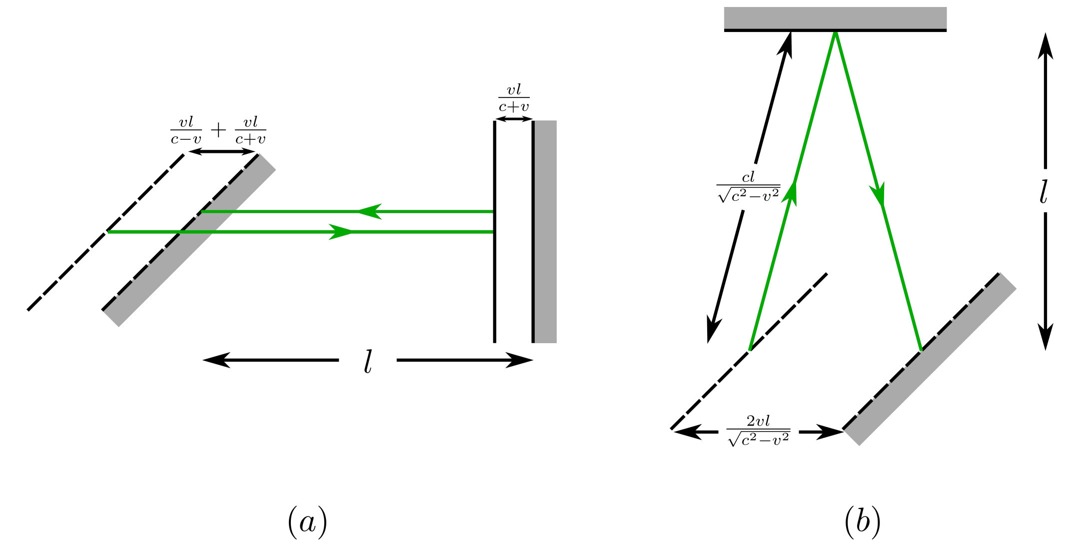


Figure 24. The geometry of the Michelson–Morley interferometer arms as viewed from the aether frame (without length contraction). Lines without a grey backing indicate the past positions of mirrors/beam splitters when they were relevant. (*a*) shows the arm parallel to the Earth's motion through the aether, while (*b*) shows the perpendicular arm.

We can calculate the predicted phase difference between the two beams by analysing the experiment in the aether frame, where the entire apparatus moves with velocity v, as shown in Fig. 24. Let us start by considering light travelling along the parallel arm. During the journey from beam splitter to mirror, the light travels at c towards the mirror, which moves away from the beam with speed v. Since, the distance between beam splitter and mirror is l, the journey therefore takes a time $l/(c-v)$, and hence the distance travelled is $lc/(c-v)$. When the light is reflected back the way it came, the beam splitter is now moving towards the beam, meaning the journey takes a time $l/(c+v)$, and thus a distance $lc/(c+v)$. As such, the total phase accumulated by the beam is

$$\phi_{\parallel} = \frac{l}{\lambda}\left(\frac{c}{c-v} + \frac{c}{c+v}\right) = \frac{2l}{\lambda} + \frac{2lv^2}{\lambda c^2} + \mathcal{O}(lv^4/\lambda c^4)\,. \tag{5.11}$$

In the aether frame, the light traversing the perpendicular arm traces out an isosceles triangle. The height of this triangle is l, and the sine of the base angle is v/c. Putting this together, we can deduce that the long sides of the triangle have lengths of $cl/\sqrt{c^2-v^2}$. This implies that the perpendicular beam accumulates a total phase

$$\phi_{\perp} = \frac{2l}{\lambda}\frac{c}{\sqrt{c^2-v^2}} = \frac{2l}{\lambda} + \frac{lv^2}{\lambda c^2} + \mathcal{O}(lv^4/\lambda c^4)\,. \tag{5.12}$$

Since the rest of the beams' journeys are along the same path, the total phase difference between them must be given by

$$\Delta\phi = \phi_{\parallel} - \phi_{\perp} = \frac{lv^2}{\lambda c^2} + \mathcal{O}(lv^4/\lambda c^4)\,. \tag{5.13}$$

For visible light, which has a wavelength $\sim 500\,\mathrm{nm}$, and using the orbital velocity of the Earth $v/c \approx 10^{-4}$, a measurable phase difference should be possible with a reasonable size of interferometer.[2] However, when Michelson and Morley actually conducted the experiment, they observed no phase difference, despite an experimental error that was forty times smaller than the predicted size of the effect. The only conclusion they could reach was that the aether must be completely dragged along with the Earth, just as Stokes had predicted.

From a special relativistic point of view, it is not at all surprising that the Michelson–Morley experiment measures no phase difference. After all, since the speed of light is constant in all inertial frames, the beams must both travel at c in the Earth frame, where the lengths of their paths are identical by construction, and so could not possibly have a phase difference. Of course, we should also be able to carry out the analysis in the Sun's rest frame, so there must be a relativistic correction to the 'aether frame' analysis we have presented. The issue is that in our prior derivation we did not account for the length contraction of the parallel interferometer arm. This length contraction precisely cancels out the contribution due to the motion of the apparatus, so the relativistic prediction is definitely no phase difference between the beams.

Individually none of the three experiments we have just discussed are fantastically strong evidence for special relativity. This is because in each case either Fresnel's or Stokes' aether model is also capable of explaining the results. It is only when we take the experiments together that we see no single aether theory (other than Lorentz's which is just a more confusing version of special relativity) can explain them all, while special relativity has no such problems.

5.4 Atmospheric muons

While the experiments we have discussed so far were historically important in establishing the need for special relativity, no skeptic is going to be convinced by them alone. It would be much more compelling if we could demonstrate experimental observations of some of the key phenomena predicted by the special theory. Of these predictions, time dilation is probably the hardest to accept, so it makes sense to start there. Fortunately, there is a wealth of experimental evidence in favour of its existence.

One such piece of evidence is the detection of high energy muons at sea level. Muons are unstable particles and decay into electrons and neutrinos with a half life of about $3.2\,\mu\mathrm{s}$, measured in their rest frame. Muons are created on Earth when high energy cosmic rays collide with atoms in the upper atmosphere, about $15\,\mathrm{km}$ above sea level. These muons move at close the speed of light, so it takes them about $50\,\mu\mathrm{s}$ to reach the ground. Without relativistic effects, we would only expect one in every $2^{50/3.2} \approx 50,000$ muons to reach sea level without decaying. If this were actually the case we would expect muon detections to be significantly outnumbered by high

[2] For a significant phase difference l would need to be at least $10\,\mathrm{m}$ or so. This is a bit large to be practical (although certainly achievable) but one can circumvent the issue by reflecting the beams multiple times before recombination to effectively increase the path lengths.

energy electrons from their decays. In reality, muons are the most common energetic particles detected on the Earth's surface, suggesting that the fast moving particles do indeed experience decay more slowly than they do at rest.

The Rossi–Hall and Frisch–Smith experiments explored this phenomenon quantitatively by measuring the energies and numbers of muons at two different heights. They were able to confirm that the decay rate of the muons measured in the Earth's reference frame was consistent with the predicted relativistic time dilation. These experiments are particularly important because muon decays occur via the weak nuclear force, and so the fact that they are affected by time dilation is an important piece of evidence that the principle of relativity applies to all physical processes, and not just electromagnetic ones.

5.5 The Hafele–Keating experiment

In 1971, Joseph Hafele and Richard Keating took a series of caesium atomic clocks onto commercial airliners to measure their time dilation. They carried out this experiment on two separate flights around the world, one travelling eastward and one travelling westward. Four clocks were taken on each, and upon landing compared the measured times to those recorded by clocks which had remained in a laboratory on the ground. By averaging the readings from the clocks they were able to bring their experimental errors down to about 10%, enough to compare against the predictions of relativity.

There is one slight subtlety to be aware of in analysing this experiment and that is the Earth's rotation. This means that the surface of the Earth is not an inertial reference frame, and so we cannot apply our usual formulae to velocities measured relative to the Earth. Instead, we work in the inertial frame where the Earth's centre of mass is at rest. If an object is located at latitude λ, and moves on a bearing θ with a speed u relative to the Earth's surface, then its speed in this inertial frame is given by

$$v^2 = u^2 + 2uR\Omega\cos\lambda\sin\theta + R^2\Omega^2\cos^2\lambda\,, \tag{5.14}$$

where $R \approx 6400\,\mathrm{km}$ is the radius of the Earth, and $\Omega \approx 7.3\times10^{-5}\,\mathrm{rad\,s^{-1}}$ is the angular frequency of the Earth's rotation. Since all speeds are much slower than light, we can safely approximate

$$\tau = (1 - v^2/2c^2)t\,, \tag{5.15}$$

where τ is the proper time measured by the moving clock, and t is the time measured by a stationary clock. If τ_+, τ_- and τ_0 are the times measured by equatorial ($\lambda = 0$) east moving ($\theta = \pi/2$), west moving ($\theta = 3\pi/2$) and stationary clocks respectively, then to the same order of approximation

$$\tau_0 - \tau_\pm = (u^2 \pm 2uR\Omega)\tau_0/2c^2\,. \tag{5.16}$$

Hafele and Keating found their results were in agreement with this formula.[3]

[3] Technically they had to numerically integrate variations in speed latitude etc. over the course of the flight, plus an altitude dependent general relativistic effect, but the point still stands.

5.6 Bucherer's experiment

While perhaps not quite on the level time dilation, special relativity's departure from the Newtonian expression for momentum is still a rather extreme prediction. One of the first experiments to measure the dependence of momentum on velocity was carried out in 1908 by Alfred Bucherer. This experiment relies on the behaviour of charged particles in electromagnetic fields. They key point is that, so long as we interpret force as the time derivative of momentum, the standard expressions for electric and magnetic forces apply even when particles are moving at relativistic speeds. This is certainly a non-trivial assumption; however, considerations based on Maxwell's equations, which are beyond the scope of our discussion, make it seem the most plausible option.

Bucherer used salts of radium (which undergoes beta decay) as a source of electrons with a range of relativistic energies. In order to control their velocities, he passed the electrons through perpendicular electric and magnetic fields, as shown in Fig. 25. The electrons entered the velocity selector with perpendicular to the crossed fields, meaning they experienced an electric force eE, and an opposite magnetic force of evB. Thus, for an electron to be undeflected by the fields it must have a velocity

$$v = \frac{E}{B} \,. \tag{5.17}$$

These electrons were then allowed to exit the velocity selector, after which they moved solely under the influence of the magnetic field. In this region they followed circular arcs with a radius given by

$$R = \frac{p}{eB} \,. \tag{5.18}$$

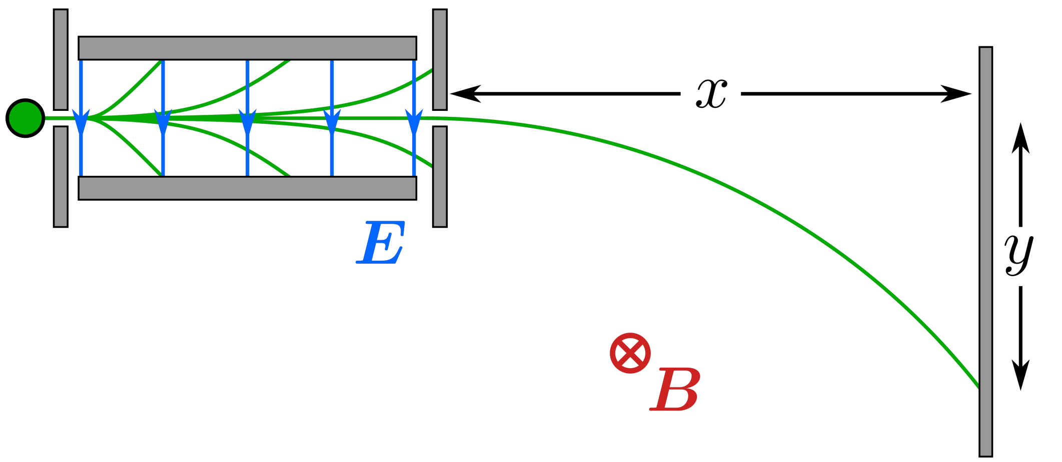


Figure 25. The basic experimental setup used by Bucherer to determine the relationship between an electron's velocity and its momentum. Electrons produced by beta decay of radium are passed through a velocity selector formed from a parallel plate capacitor and a uniform magnetic field. Electrons which are undeflected are then allowed to move under the influence of the magnetic field until they reach a photographic plate.

This expression derives from the fact that the magnetic force is evB, while the rate of change of the electron's momentum will be vp/R, both perpendicular to the magnetic field and the electron's velocity. The electrons were detected by a photographic film a distance x from the velocity selector. By noting that the deflection of the electron on the plate must obey

$$x^2 + (R - y)^2 = R^2 \implies x^2 + y^2 - 2Ry = 0\,, \tag{5.19}$$

we can infer that the radius of the electron's path was given by

$$R = \frac{x^2 + y^2}{2y}\,. \tag{5.20}$$

By varying and recording the values of E and B, this apparatus provides a way to measure the velocity-momentum relationship of electrons with speeds up to $0.9c$. With the results of his experiment, and a subsequent refinement carried out by his student Kurt Wolz in 1909, Bucherer was able to confirm that the momentum of an electron was indeed given by

$$p = \frac{mv}{\sqrt{1 - v^2/c^2}}\,, \tag{5.21}$$

just as special relativity says it should be.

5.7 The Cockcroft–Walton experiment

The last major prediction of special relativity is the relationship between an object's mass and its energy content. In 1932, John Cockcroft and Ernest Walton were the first to provide experimental evidence of the equivalence between mass and energy. They studied the nuclear reaction

$${}^1_1\mathrm{H} + {}^7_4\mathrm{Li} \to {}^4_2\mathrm{He} + {}^4_2\mathrm{He}\,, \tag{5.22}$$

by accelerating protons up through hundreds of kilovolts into a stationary lithium target. They determined the kinetic energies of the alpha particles by measuring the thickness of mica plate required to absorb them. They found that the alpha particles carried a total kinetic energy of roughly 17 MeV split evenly between them. Importantly, all of the alpha particles had the same energy, which suggests that they are the only products of (5.22), and no energy is lost to gamma rays as is common in nuclear reactions.

The masses of the nuclei were sufficiently well known at the time to deduce that the reaction involved a mass decrease of 0.0154 ± 0.003 Da (0.0186 Da using modern data).[4] Multiplying by the speed of light squared, this mass difference should translate into a release of 14.3 ± 2.7 MeV (17.3 MeV using modern data) of energy, which is consistent with the observations of the alpha particles. Not only do these results help verify special relativity, but, because nuclear reactions are primarily governed by the strong nuclear force, they also help to show that relativity applies to all physical interactions and not just electromagnetism.

[4] The Dalton is a unit of mass equal to one twelfth the mass of a carbon-12 atom. $1\,\mathrm{Da} = 1.660\,539\,066\,60(50) \times 10^{-27}\,\mathrm{kg} = 931.49410242(28)\,\mathrm{MeV}/c^2$.

6 Relativity as Spacetime Geometry

If we were to make a complaint about the version of special relativity presented so far, it would probably be that it is algebraically inelegant. Having to remember different transformation laws for positions, velocities, frequencies, etc. can be tiresome, a problem which only gets worse if we were to introduce things like electromagnetic fields, which have yet another transformation law. Even when the same laws appear in multiple places, such as the reappearance of Lorentz transformations for energy and momentum, we have no real insight into why this is the case. The theory as a whole is crying out for these transformations to be combined into a single unified framework. Fortunately, such a framework was discovered in 1907 by Hermann Minkowski, who realised that special relativity takes a particularly simple form when regarded as a statement about the geometry of four dimensional spacetime.

The central idea behind this approach is that we can draw a number of equivalences between concepts in relativistic physics and those that arise in elementary geometry. In particular, we can identify inertial paths through spacetime with the notion of a straight line in standard Euclidean geometry. This, in turn, gives rise to the identification of time measured by a clock along its path through spacetime and the distance measured along a curve. The principle of relativity gains a particularly simple interpretation as the analogue of rotational symmetry in more standard spatial geometry. It is no secret that ordinary geometry can be greatly simplified via the introduction of vectors and their associated notation. Perhaps unsurprisingly, the same is true of spacetime geometry. These spacetime vectors, usually referred to as 4-vectors because they are four dimensional, form the backbone of the geometric approach to special relativity.

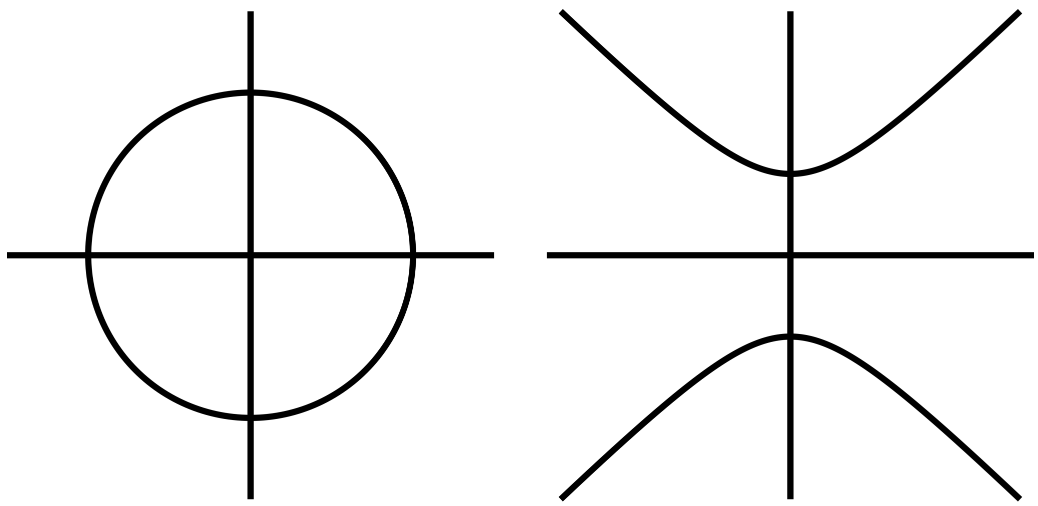


Figure 26. On the left is a circle as described by standard two dimensional Euclidean geometry. On the right is the equivalent construction in a two dimensional slice of Minkowski's spacetime geometry. The curve consists of all events which can be reached in a fixed amount of proper time by an inertial observer starting at the origin.

6.1 A review of Euclidean geometry

Before we dive into the details of Minkowski's spacetime geometry, it will be worthwhile for us to recap some basic Euclidean geometry, to get ourselves used to thinking geometrically. The defining features of Euclidean geometry are the transformations one can carry out which leave the distances between points unchanged. These are: translations, rotations, and reflections.[1] Objects which can be mapped onto one another by some combination of these transformations are referred to as congruent, as shown in Fig. 27. Closely related to congruency is the notion of similarity. Two objects are similar to one another if the symmetry transformations allow one to be mapped into a rescaled copy of the other. That is to say, given a pair of similar objects, the ratio between corresponding lengths on each object is a fixed constant. Importantly, since angles measure ratios between lengths, two objects will be similar if and only if all their angles are the same.

There are a number of useful theorems, such as results about sums of angles in different polygons, or relationships between angles when lines intersect one another, which can be used to help establish either similarity or congruence. We shall not discuss the proofs and derivations of these results in any great detail, and shall instead simply note that they all follow from careful applications of the previously mentioned symmetries.

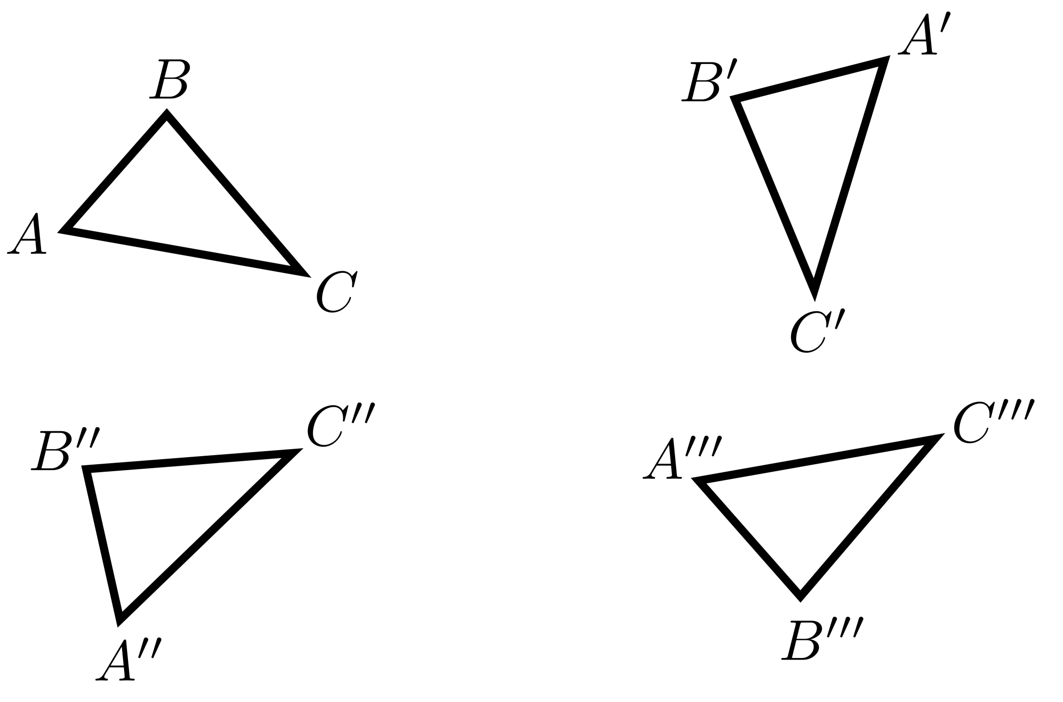


Figure 27. Four congruent triangles. Since each one can be related to every other by symmetries of the underlying Euclidean geometry, any geometrical properties such as side lengths and angles must be the same for each triangle.

[1] An important property of Euclidean geometry is that translations in different directions commute with one another, unlike in the similarly symmetric hyperbolic geometry.

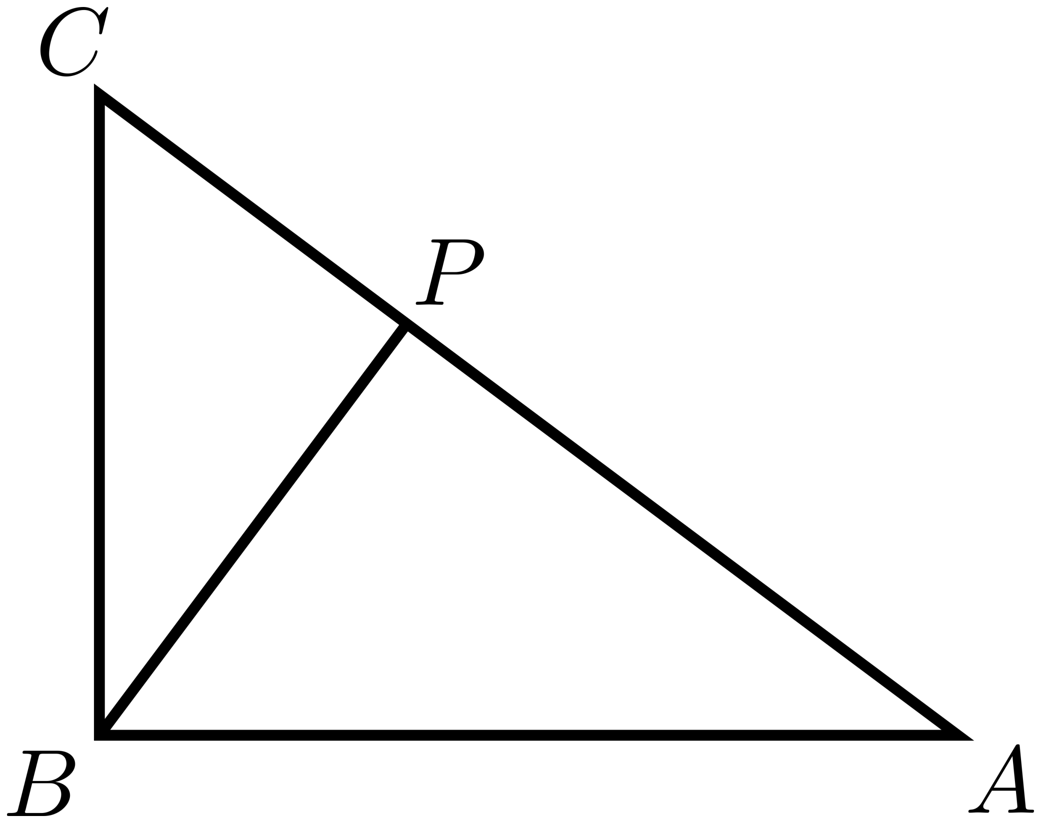


Figure 28. The construction used in Einstein's proof of Pythagoras' theorem. We start with a right angled triangle ABC. The point P is chosen along the line segment AC such that BP is perpendicular to AC. This splits the original triangle into two smaller right angled triangles: BPC and APB. In addition to the right angle, each of these triangles has a second angle which is the same as the original, and, since the sum of the angles in a triangle is fixed, thus they are both similar to ABC.

One of the most important results in Euclidean geometry, which we shall discuss in some detail, is Pythagoras' theorem. There a great many different methods of proving this theorem to choose from; however, given the subject matter at hand it seems only fitting for us to look at one that was reportedly discovered by Einstein himself while he was in school. The basic geometric construction used in this proof is shown in Fig. 28. The idea is that, because a right angle is exactly half of a straight line, we can split the triangle into two smaller right angled triangles which are similar to the original. Using the labelling of points from the figure, we can say that the length of the hypotenuse is given by

$$\overline{AC} = \overline{AP} + \overline{PC}\,. \tag{6.1}$$

To determine the lengths of the two line segments, we can multiply the length of the corresponding side in the original triangle by the scale factor between the two similar triangle. This scale factor is most conveniently expressed as the ratio between the hypotenuse of each triangle, which gives us

$$\overline{AP} = \overline{AB}\,\frac{\overline{AB}}{\overline{AC}} \quad \text{and} \quad \overline{PC} = \overline{BC}\,\frac{\overline{BC}}{\overline{AC}}\,. \tag{6.2}$$

Substituting these expressions into (6.1), and multiplying by $\overline{AC}$, directly gives us Pythagoras' theorem

$$\left(\overline{AC}\right)^2 = \left(\overline{AB}\right)^2 + \left(\overline{BC}\right)^2 . \tag{6.3}$$

It is often convenient for us to describe positions on the plane by assigning each point a unique numerical label referred to as its coordinates. There are numerous ways of doing this, but the simplest and most common is the Cartesian system. The first step is to identify a particular special point and label it as the origin O. We then construct a set of coordinate axes, which is a pair of perpendicular lines passing through O. The process of using these axes to assign coordinates to a point A is shown in Fig. 29. The idea is to construct a pair of lines through A that are parallel to the coordinate axes. Together with the axes themselves, these lines form a rectangle, and the point's coordinates (x_A, y_A) are the signed distances along of the sides of this rectangle.

This coordinate system is particularly advantageous because it makes calculating distances straightforward. Given any two points in the plane with coordinates (x, y) and $(x + \Delta x, y + \Delta y)$, we can draw a right angled triangle such that the hypotenuse connects the two points, and the other two sides are parallel to the axes. The lengths of these sides will be Δx and Δy, so Pythagoras' theorem tells us that the distance between them is

$$(\Delta l)^2 = (\Delta x)^2 + (\Delta y)^2 . \tag{6.4}$$

Naturally, there are many different equally valid choices of Cartesian coordinates that can be obtained via different choices of origin, or different orientations of the axes. The distances between points should be independent of the way we choose to describe them, and the reasoning behind (6.4) holds true for all sets of Cartesian coordinates. Thus, if x, y and x', y' are two sets of Cartesian coordinates, they must be related in such a way that, for any pair of points

$$(\Delta x)^2 + (\Delta y)^2 = (\Delta x')^2 + (\Delta y')^2 . \tag{6.5}$$

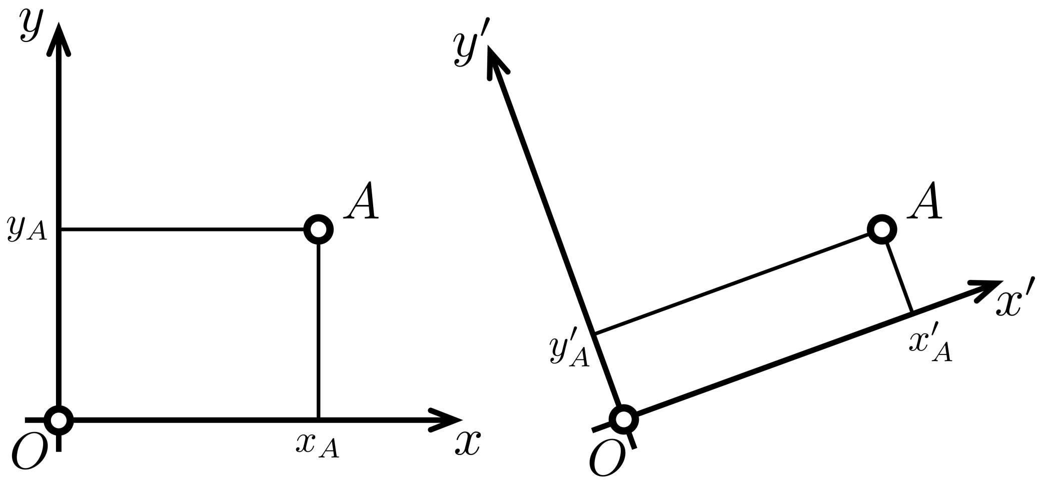


Figure 29. The coordinates of a point A determined with respect to two different sets of coordinate axes. Since the origin is the same point for both sets of axes, and the distance between that point and A is independent of the coordinates we use, we must have $x_A^2 + y_A^2 = x_A'^2 + y_A'^2$.

This result is trivially true for translations of the origin, which just change the coordinates of every point by a constant offset, and hence have no effect on coordinate differences. Similarly, a reflection that simply flips the sign of one coordinate will do nothing to the squared differences, and so will easily satisfy (6.5). If we keep the origin fixed, but rotate the axes through an angle θ, then we can do some trigonometry to deduce that the new coordinates are given by

$$x' = x\cos\theta + y\sin\theta \quad \text{and} \quad y' = y\cos\theta - x\sin\theta\,. \tag{6.6}$$

Direct substitution allows us to confirm that this too satisfies (6.5). The interesting thing is that the most general transformation possible which obeys (6.5) can be represented as the composition of translation, reflection, and rotation. In other words, a necessary and sufficient condition for a set of coordinates to be Cartesian is that distances are given by (6.4).

We shall conclude our discussion of Euclidean geometry with a brief overview of vectors. There is a wealth of formal mathematics involved in precisely defining a vector; however, for our purposes it will be sufficient for us to imagine a vector as an arrow in space. This arrow has two key properties: magnitude (i.e the length of the arrow) and direction (i.e which way the arrow is pointing). We can use coordinate axes to split a vector into components, which we denote as

$$\boldsymbol{a} = (a_x, a_y)\,, \tag{6.7}$$

where the components a_x and a_y indicate the coordinate differences between the head and tail of the vector arrow. Importantly, this definition means that the components of a vector obey the same transformation laws as the coordinates under reflections and rotations, while they are invariant under translations. Using Pythagoras' theorem, the magnitude of a vector is given by

$$|\boldsymbol{a}|^2 = a_x^2 = a_y^2\,, \tag{6.8}$$

which will be invariant under coordinate transformations in the same way that distances are. We can add vectors together using the intuitive formula

$$\boldsymbol{a} + \boldsymbol{b} = (a_x + b_x, a_y + b_y)\,. \tag{6.9}$$

Together with the concept of magnitude, this enables us to construct an invariant product between two vectors, often called the dot or scalar product, defined as

$$\boldsymbol{a} \cdot \boldsymbol{b} = \frac{|\boldsymbol{a} + \boldsymbol{b}|^2 - |\boldsymbol{a} - \boldsymbol{b}|^2}{4} = a_x b_x + a_y b_y\,. \tag{6.10}$$

The value of the dot product lies in the fact that we get the same answer no matter which set of coordinate axes we use to evaluate it. So for example, suppose we want to determine the angle θ between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. We know that, if we were to choose a set of coordinate axes such that the x axis was parallel to $\boldsymbol{a}$, then the two vectors would have components $(|\boldsymbol{a}|, 0)$ and $(|\boldsymbol{b}|\cos\theta, |\boldsymbol{b}|\sin\theta)$. In principle, we could determine the appropriate rotation to bring the components of $\boldsymbol{a}$ into this form, apply that transformation to the components of $\boldsymbol{b}$, and then read of the angle. However, a much quicker and easier way is to note that we have

$$\cos\theta = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\sqrt{(\boldsymbol{a} \cdot \boldsymbol{a})(\boldsymbol{b} \cdot \boldsymbol{b})}}\,, \tag{6.11}$$

and since the dot product has the same value in any coordinate system, we can evaluate the right hand side using whatever components are readily available.

6.2 Minkowski geometry

We can build up the four dimensional Minkowski geometry of spacetime in much the same way as we did for the Euclidean plane. However, before we begin, we need to first establish what exactly 'distance' is in the context of spacetime geometry. We call two events timelike separated if the straight line joining them would represent an object moving slower than light. If A and B aretimelike separated events, we define the distance between them $\overline{AB}$ to be c multiplied by the time measured by an inertial clock passing between the two events. The factor of c is included simply to make sure that this distance has the appropriate units.

If an object travelling between two events would have to move faster than light, we refer to the events as being spacelike separated. If A and B are two such events, we define the distance between them as follows. Firstly, we construct an inertial worldline through A, and define P and Q as the events along that line which can be connected to B via light rays. We then adjust the worldline until $\overline{AP} = \overline{AQ}$. A line with this property is said to be orthogonal to AB, which is the Minkowski analogue of being perpendicular. Once we have found an orthogonal line, we simply make the definition $\overline{AB} = \overline{AP} = \overline{AQ}$. The motivation behind this procedure is that an observer whose worldline goes through A can measure the distance to B by timing how long it takes for light to reach B and be reflected back. The orthogonality of PQ and AB is necessary to ensure that the observer regards A and B as simultaneous.

Just like its Euclidean counterpart, Minkowski geometry is fundamentally defined by its symmetries. Given the context of this discussion, it should not be especially surprising that the symmetries of Minkowski spacetime are identical to the transformations which convert between inertial reference frames. In practice, this means that two collections of events Γ and Γ' will be congruent if and only if we can find two inertial observers $\mathcal{O}$ and $\mathcal{O}'$ such that $\mathcal{O}$'s measurements of Γ are identical to

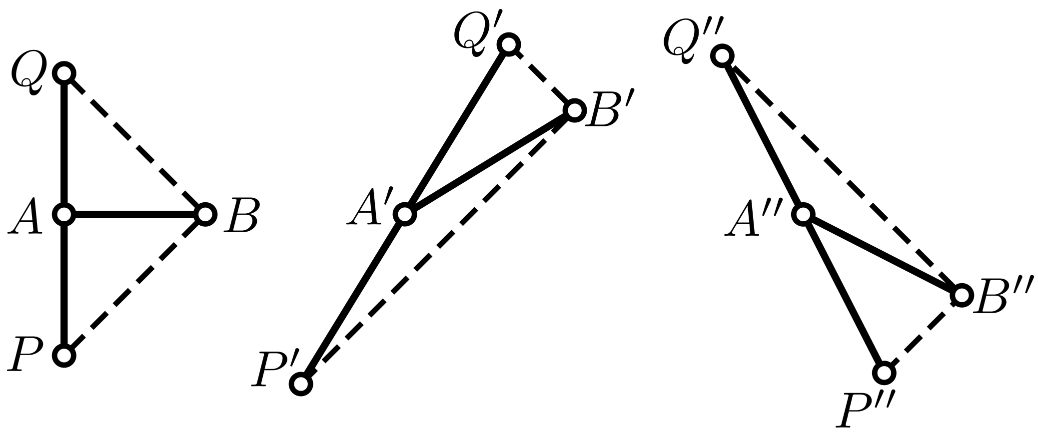


Figure 30. Three different pairs of orthogonal lines in Minkowski spacetime. The dashed lines represent the paths of light rays. If we imagine three inertial observers $\mathcal{O}, \mathcal{O}', \mathcal{O}''$, whose worldlines lie along $PQ, P'Q', P''Q''$ respectively, then each observer's measurements of their appropriately primed events will be the identical. Thus, the three collections of events are congruent to one another.

$\mathcal{O}'$'s measurements of Γ'. This gives rise to three distinct types of transformation: spacetime translations, spatial rotations, and boosts. The translations and rotations are not particularly interesting; they simply ensure that, when an observer identifies a slice of spacetime containing simultaneous events, that slice has three dimensional Euclidean geometry. It is the boosts that give Minkowski geometry its unique character, and we shall use them to derive an analogue of Pythagoras' theorem.

Let us first consider a pair of timelike separated events A and C. We start by constructing an arbitrary inertial worldline that passes through A, and define P and Q as the events along that worldline that can be connected to C via light rays. From here, the key insight is to realise that the triangles ACQ and APC are similar. This can be seen in Fig. 31, where APC is boosted to make AC parallel to AQ, and then rotated in space so that AP is parallel to AC. Once we have established this similarity, we know that the ration between corresponding lengths must be a constant. Hence, we can deduce that

$$\frac{\overline{AC}}{\overline{AQ}} = \frac{\overline{AP}}{\overline{AC}} \implies \left(\overline{AC}\right)^2 = \left(\overline{AP}\right)\left(\overline{AQ}\right). \tag{6.12}$$

Let us now suppose that A and C are spacelike separated. We can obtain essentially the same result through very similar reasoning. Just as before, we imagine an inertial worldline through A, and let P and Q denote the events on that worldline which can be reached from C by light rays. This time A will be between P and Q as opposed to on one side of them. Next, we construct a line through A that is orthogonal to PQ, and lying in the plane defined by PQ and AC, as shown in Fig. 32. We now

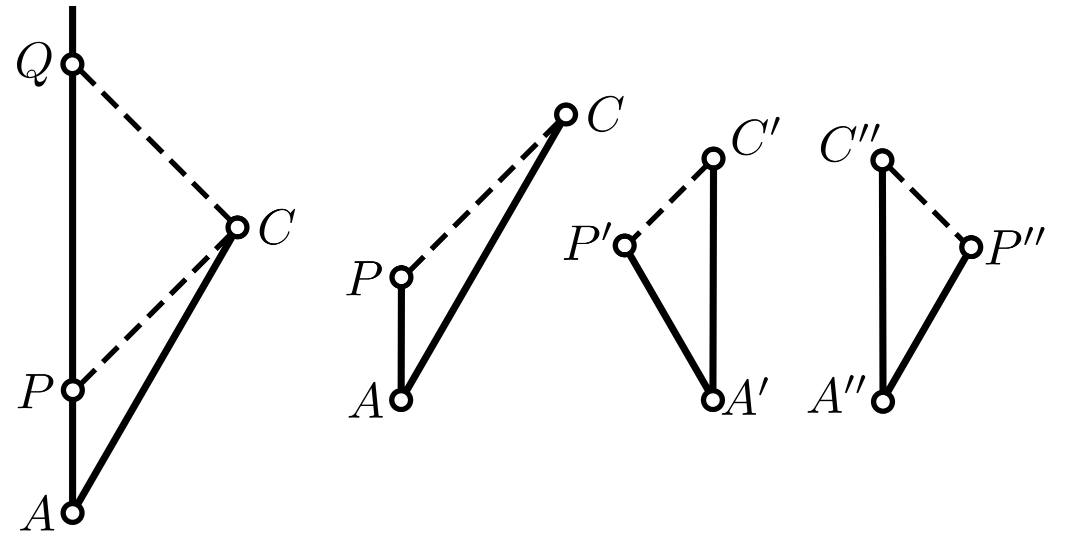


Figure 31. The construction used to determine the length of timelike line segment AC. The key argument is that triangles APC and ACQ are similar. This is shown by the triangles $A'P'C'$ and $A''P''C''$ which are congruent to APC. $A'P'C'$ is obtained from ABC via a boost, and $A''P''C''$ by applying a further rotation about $A'C'$. The key point is that after these symmetry transformations $P''C''$ is still a light ray. This is the geometrical origin of the light postulate.

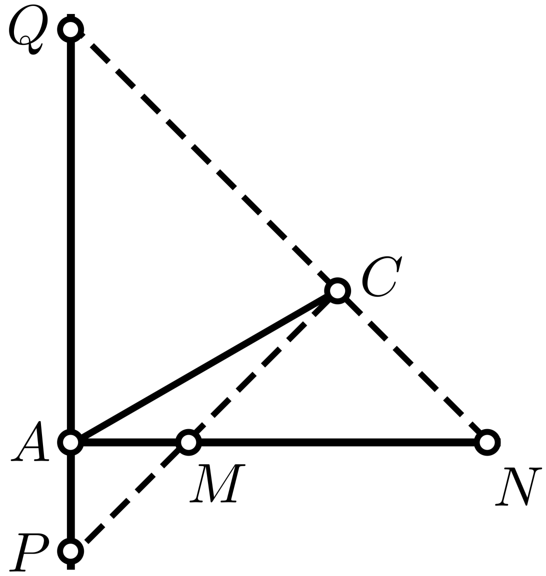


Figure 32. The construction used to determine the length of a spacelike line segment AC. Note that, by definition, if a triangle is formed by two orthogonal lines and a light ray, the two orthogonal lines must have the same length. Thus in this construction $\overline{AM} = \overline{AP}$ and $\overline{AN} = \overline{AQ}$.

define two additional events: M and N, which are the events along this line that can be connected to C using light rays.

Following an identical argument to the timelike case, the triangles AMC and ACN are similar to one another, and so we can say that

$$\frac{\overline{AC}}{\overline{AN}} = \frac{\overline{AM}}{\overline{AC}} \implies \left(\overline{AC}\right)^2 = \left(\overline{AM}\right)\left(\overline{AN}\right). \tag{6.13}$$

Looking at Fig. 32, we can see that from the definitions of orthogonality and the length of spacelike lines that we must have $\overline{AP} = \overline{AM}$ and $\overline{AQ} = \overline{AN}$. Thus, we once again conclude that

$$\left(\overline{AC}\right)^2 = \left(\overline{AP}\right)\left(\overline{AQ}\right). \tag{6.14}$$

To obtain an equivalent of Pythagoras' theorem, we will need to consider something equivalent to a right angled triangle. Thus, let us consider yet another event B which lies along the inertial worldline in such a place that the line BC is orthogonal to AB, as shown in Fig. 33. Once we have done this, we can use the definition of orthogonality to deduce that

$$\overline{AQ} = \overline{AB} + \overline{BC} \quad \text{and} \quad \overline{AP} = \pm\left(\overline{AB} - \overline{BC}\right), \tag{6.15}$$

where the sign of the $\pm$ is determined by whether the events are timelike or spacelike separated. So that we can deal with this sign issue in a convenient way, we shall define a new quantity which incorporates the sign into itself. This quantity is the interval, and it is defined by

$$(\Delta s_{AC})^2 = \begin{cases} \left(\overline{AC}\right)^2 & AC \text{ timelike separated} \\ -\left(\overline{AC}\right)^2 & AC \text{ spacelike separated} \end{cases}. \tag{6.16}$$

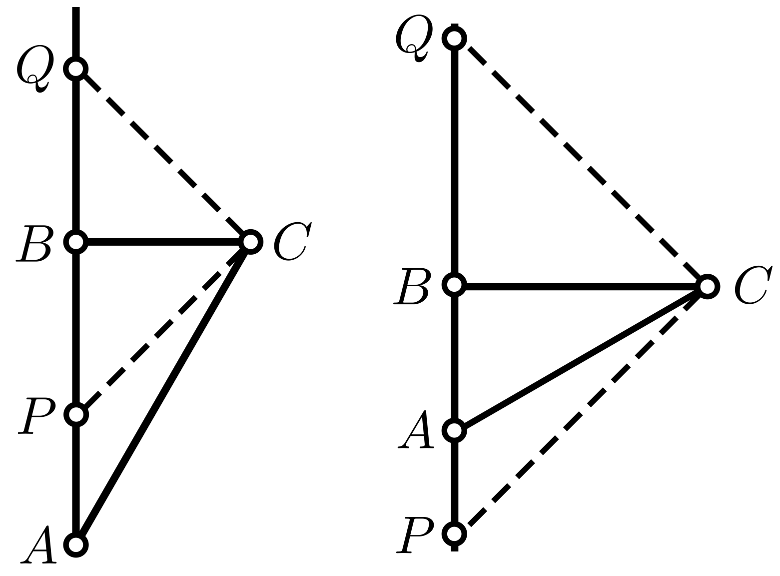


Figure 33. The Minkowski equivalents of right angled triangles.

Putting this all together, our final result takes the form

$$(\Delta s_{AC})^2 = \left(\overline{AB}\right)^2 - \left(\overline{BC}\right)^2 . \tag{6.17}$$

6.3 Lorentz transformations and 4-vectors

Cartesian coordinates are an invaluable tool in the study of Euclidean geometry, and so it seems sensible to attempt an analogous construction in Minkowski spacetime. We start by identifying a particular event to be the origin of our coordinate system. We then construct an inertial worldline through that event to serve as our first coordinate axis. We now consider the slice of spacetime through the origin formed of all the lines orthogonal to this axis. This slice has a three dimensional Euclidean geometry, and so we can construct our remaining coordinate axes using the same procedure as we would for Cartesian axes. We can assign coordinates to any event by constructing a hyperrectangle (4D version of a rectangle) with the event and the origin as opposite vertices, and all its edges parallel to the coordinate axes. The signed lengths of these edges w, x, y, z are referred to as the Lorentzian coordinates of the event.

By applying Pythagoras' theorem in the Euclidean space orthogonal to the w axis, and then using our expression for the interval, we can deduce that, if two events have Lorentzian coordinates (w, x, y, z) and $(w + \Delta w, x + \Delta x, y + \Delta y, z + \Delta z)$, the interval between them is given by

$$(\Delta s)^2 = (\Delta w)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 . \tag{6.18}$$

Lorentzian coordinates are nice, but, for historical reasons, it is more practical to denote an event's position along the w axis with the coordinate $t = w/c$, so that it can be measured in units of time and not distance. Expressed in these terms, the interval becomes

$$(\Delta s)^2 = (c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 \,. \tag{6.19}$$

The coordinates t, x, y, z are precisely the coordinates of an inertial frame $\mathcal{K}$ in which the w axis is at rest. Since the interval depends only on the underlying spacetime geometry, and not the coordinates we use to describe it, this implies that for any two inertial frames $\mathcal{K}$ and $\mathcal{K}'$ we must have

$$(c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 = (c\Delta t')^2 - (\Delta x')^2 - (\Delta y')^2 - (\Delta z')^2 \tag{6.20}$$

Translations, which do not affect coordinate differences, and reflections, which only change the signs of the differences, will trivially satisfy (6.20). Less simply, rotations, which preserve Euclidean distance, in the three spatial coordinates such as

$$y' = y\cos\theta - z\sin\theta \quad \text{and} \quad z' = z\cos\theta + y\sin\theta \,, \tag{6.21}$$

will also obey (6.20). Most importantly to us, Lorentz boosts which mix spatial and time coordinates, for example

$$ct' = ct\cosh\chi - x\sinh\chi \quad \text{and} \quad x' = x\cosh\chi - ct\sinh\chi \,, \tag{6.22}$$

also leave the expression for the interval unchanged. Thus, we can see that the Lorentz transformations which we previously derived in a rather algebraic manner, arise as naturally from Minkowski geometry as rotations do from Euclidean.

One of our key complaints about the algebraic relativity was that there are too many different transformation laws. It turns out that the way of avoiding this complication is to introduce the concept of a vector into spacetime. These vectors are often referred to as 4-vectors so they will not get confused with three dimensional vectors like velocity or the electric field which we are more familiar with. We can imagine a 4-vector as an arrow in 4-dimensional spacetime and, given a set of Lorentzian coordinate axes, we can assign it components based on the coordinate differences between the arrow's head and tail. When doing this we use the value of the w coordinate not the t coordinate, because this ensures that all of the components have the same units. For example, the 4-vector which represents an arrow pointing from the origin to the event A is

$$\underline{r}_A = (ct_A, x_A, y_A, z_A) \,. \tag{6.23}$$

Importantly, the components of a 4-vector will always obey the Lorentz transformations under changes of reference frame, since they are defined directly with respect to the coordinate axes. Thus, if we want to know how a particular quantity transforms, we simply need to relate it to the components of a 4-vector.

We can define a scalar product between 4-vectors as

$$\underline{r}_A \cdot \underline{r}_B = c^2 t_A t_B - x_A x_B - y_A y_B - z_A z_B \,, \tag{6.24}$$

which will be invariant under coordinate transformations for the same reasons as the Euclidean dot product.

6.4 4-velocity and 4-momentum

Let us consider the motion of a single, not necessarily inertial, particle. This motion generates a curve in spacetime, called the particle's worldline, which consists of all events at which the particle is present. If we pick a particular event along the worldline, we can define $d\underline{r}$ to be the difference in 4-position between a second event, which is infinitesimally further along the curve,and the first. Any 4-vector which is tangent to the worldline at this event must be parallel to $d\underline{r}$, and hence must be equal to $d\underline{r}$ multiplied by some scale factor. We define the particle's 4-velocity $\underline{u}$ to be the tangent vector that is scaled such that $\underline{u} \cdot \underline{u} = c^2$. Thus, the 4-velocity is given by

$$\underline{u} = \frac{c d\underline{r}}{\sqrt{d\underline{r} \cdot d\underline{r}}} = \frac{d\underline{r}}{d\tau}\,, \tag{6.25}$$

where we have noted that the proper time measured by the particle between the two events $d\tau$ is, by definition, the length of the vector connecting them divided by c, and hence $d\underline{r} \cdot d\underline{r} = c^2 d\tau^2$.

Expanding this expression in the coordinates of an inertial frame $\mathcal{K}$, we find that

$$\underline{u} = \left(\frac{d(ct)}{d\tau}, \frac{dx}{d\tau}, \frac{dy}{d\tau}, \frac{dz}{d\tau}\right) = \left(\frac{dt}{d\tau}c, \frac{dt}{d\tau}\frac{dx}{dt}, \frac{dt}{d\tau}\frac{dy}{dt}, \frac{dt}{d\tau}\frac{dz}{dt}\right)\,. \tag{6.26}$$

The motivation behind using the chain rule to express $\underline{u}$ in terms of the derivatives $dx/dt, dy/dt, dz/dt$ is that they are the components of the particle's velocity

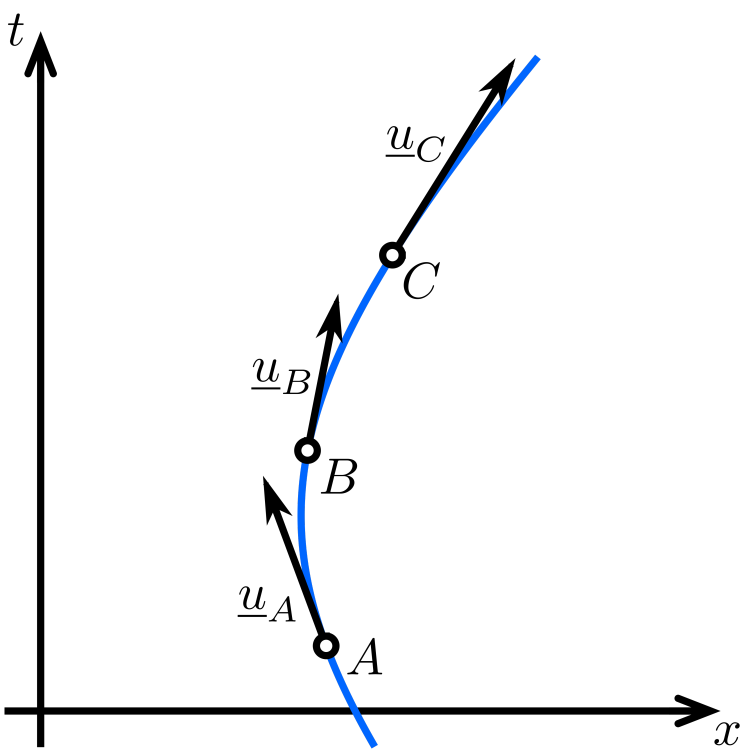


Figure 34. The blue curve represents the worldline of an accelerating particle. The particle's 4-velocity is shown at three events A, B, and C.

u_x, u_y, u_z, and so (6.26) gives us a connection between regular velocity and its 4-vector counterpart. We can find the value of $dt/d\tau$ be noting that

$$c^2 = \underline{u} \cdot \underline{u} = \left(\frac{dt}{d\tau}\right)^2 \left(c^2 - u_x^2 - u_y^2 - u_z^2\right), \tag{6.27}$$

and hence that

$$\frac{dt}{d\tau} = \frac{c}{\sqrt{c^2 - u^2}} = \gamma_u\,, \tag{6.28}$$

where $u^2 = u_x^2 + u_y^2 + u_z^2$ is the square of the particle's speed. (6.28) is simply the differential form of the integral equation we arrived at for the proper time measured by a moving clock in our original discussion of time dilation. Our final expression for the 4-velocity is thus

$$\underline{u} = (\gamma_u c, \gamma_u u_x, \gamma_u u_y, \gamma_u u_z)\,. \tag{6.29}$$

Since $\underline{u}$ is a 4-vector, we know that under a change of reference frames its components must transform in the same way as the coordinates ct, x, y, z, i.e they follow the Lorentz transformations. So, if $\mathcal{K}$ and $\mathcal{K}'$ are two frames in the standard configuration, then we must have

$$\begin{gathered}\gamma_u c = \gamma_v\left(\gamma_{u'} c + v\gamma_{u'} u_x'/c\right), \qquad \gamma_u u_x = \gamma_v\left(\gamma_{u'} u_x' + v\gamma_{u'} c/c\right), \\ \gamma_u u_y = \gamma_{u'} u_y'\,, \qquad \text{and} \qquad \gamma_u u_z = \gamma_{u'} u_z'\,.\end{gathered} \tag{6.30}$$

To isolate the velocity components, we divide the final three equations through by the first, which allows for a lot of cancellation, and eventually yields our familiar velocity transformation law

$$(u_x, u_y, u_z) = \left(\frac{u_x' + v}{1 + vu_x'/c^2}, \frac{u_y'}{\gamma_v(1 + vu_x'/c^2)}, \frac{u_z'}{\gamma_v(1 + vu_x'/c^2)}\right). \tag{6.31}$$

One point worth being aware of is that, strictly speaking, we cannot define the 4-velocity for a particle moving at the speed of light. We can see that, because γ_c involves division by zero and is hence undefined, (6.29) will fail when applied to a particle moving at c. The problem arises because the 4-velocity must have a magnitude of c, while any 4-vector tangent to a lightlike worldline has magnitude zero. Fortunately, the velocity transformation law still holds, which we could prove by directly considering the transformations of the 4-vector $d\underline{r}$ without attempting to normalise it first.

In addition to the 4-velocity, there is a second 4-vector associated with a particle's motion. This is the 4-momentum, which has components

$$\underline{p} = (E/c, p_x, p_y, p_z)\,. \tag{6.32}$$

It is not obvious a priori that this particular combination of energy and momentum will produce a 4-vector. The ultimate reason why they must do so lies in a result known as Noether's theorem. This theorem essentially states that the conservation of energy and momentum arise as consequences of the fact that the laws of physics are invariant under temporal and spatial translations. This establishes a relationship between the conserved quantities and directions in spacetime which ensures that the 4-momentum is a 4-vector. Since the magnitude of the 4-momentum is frame invariant, it must be an intrinsic property of the particle itself, and not of the particles motion. We can identify this property with the particle's mass via the relation

$$m^2 = \frac{\underline{p} \cdot \underline{p}}{c^2} = \frac{E^2}{c^4} - \frac{p_x^2 + p_y^2 + p_z^2}{c^2} \,. \tag{6.33}$$

By symmetry, the 4-momentum must be tangent to the particle's worldline, and so the mass will be zero if and only if the worldline is lightlike, meaning that the particle is travelling at the speed of light. If the mass is non-zero, then this implies that we must have

$$\underline{p} = m\underline{u} = (\gamma_u mc, \gamma_u mu_x, \gamma_u mu_y, \gamma_u mu_z) \,, \tag{6.34}$$

demonstrating that the mass obtained from the magnitude of the 4-momentum is the Newtonian mass in the low velocity limit, and reproducing our prior formulae for the relativistic energy and momentum.

To demonstrate how these ideas can apply in practice, let us once again consider the problem of Compton scattering. We shall let $\underline{p}_i, \underline{q}_i, \underline{p}_f, \underline{q}_f$ represent the relevant 4-momenta, as shown in Fig. 35. The conservation laws for energy and momentum can be brought together into the single statement of 4-momentum conservation

$$\underline{p}_i + \underline{q}_i = \underline{p}_f + \underline{q}_f \,. \tag{6.35}$$

By manipulating this equation, we can derive a number of useful results. For example, suppose we are primarily interested in the behaviour of the scattered photon. We can eliminate the scattered electron's 4-momentum by isolating it and then squaring it to obtain

$$(\underline{p}_i + \underline{q}_i - \underline{q}_f) \cdot (\underline{p}_i + \underline{q}_i - \underline{q}_f) = \underline{p}_f \cdot \underline{p}_f \,. \tag{6.36}$$

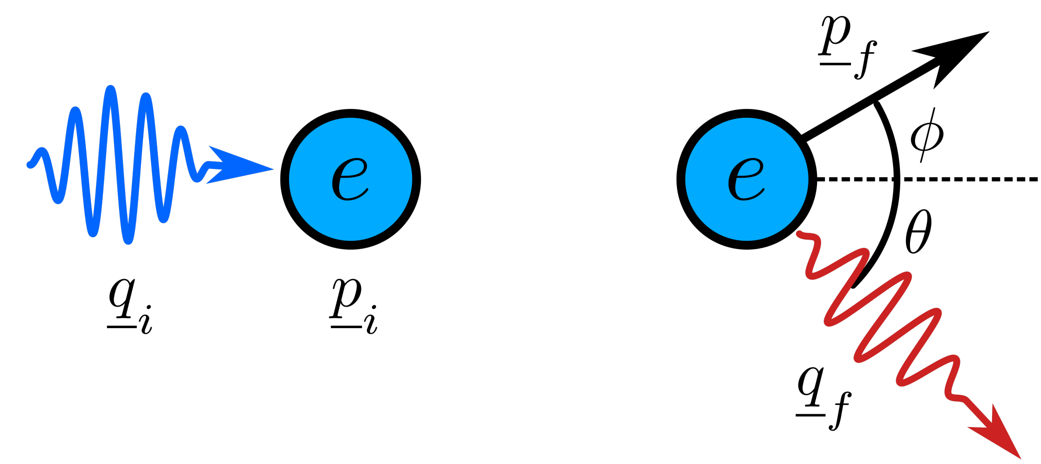


Figure 35. A photon of 4-momentum $\underline{q}_i$ scattering off an electron with 4-momentum $\underline{p}_i$. After the collision, the electron has 4-momentum $\underline{p}_f$, while the photon has $\underline{q}_f$.

Expanding out the brackets and recognising that $\underline{p}_i \cdot \underline{p}_i = \underline{p}_f \cdot \underline{p}_f = m_e^2 c^2$, while $\underline{q}_i \cdot \underline{q}_i = \underline{q}_f \cdot \underline{q}_f = 0$, we can deduce that

$$\underline{p}_i \cdot \underline{q}_i - \underline{p}_i \cdot \underline{q}_f - \underline{q}_i \cdot \underline{q}_f = 0\,. \tag{6.37}$$

This equation allows us to deduce the properties of the scattered photon, provided we know the initial conditions. If we choose an inertial frame of reference where the electron is initially at rest and orient the spatial axes such that the photon initially travels along the x direction and scatters in the x, y plane, then the 4-momenta are

$$\begin{aligned} \underline{p}_i &= (m_e c, 0, 0, 0)\,, \quad \underline{q}_i = (E_i/c, E_i/c, 0, 0)\,, \\ \text{and} \quad \underline{q}_f &= (E_f/c, (E_f/c)\cos\theta, (E_f/c)\sin\theta, 0)\,, \end{aligned} \tag{6.38}$$

where E_i and E_f are the energy of the initial and final photons respectively, and θ is the angle through which the photon is scattered. Substituting these components into (6.37), we find that

$$m_e E_i - m_e E_f - \frac{E_i E_f (1-\cos\theta)}{c^2} = 0 \implies E_f = \frac{m_e c^2 E_i}{m_e c^2 + E_i(1-\cos\theta)}\,, \tag{6.39}$$

which is the same result we derived earlier via an equivalent but slightly more algebraicly intense method.

6.5 The wave 4-vector

Taking a geometric approach to special relativity can also greatly simplify our analysis of wave phenomena. Previously, we have seen that a wave can be described by a function of the form

$$\psi(t, x, y, z) = A\cos\left(2\pi\nu t - \frac{2\pi}{\lambda}(x\cos\theta_x + y\cos\theta_y + z\cos\theta_z) + \phi\right)\,, \tag{6.40}$$

and we used the fact that the argument of the cosine must be invariant between different reference frames to deduce complicated transformation laws for the frequency ν and wavelength λ. The geometric 4-vector approach gives us a much more elegant way of approaching this problem. First of all, we define the wave 4-vector $\underline{k}$ whose components in an inertial frame $\mathcal{K}$ are given by

$$\underline{k} = (2\pi\nu/c, 2\pi\cos\theta_x/\lambda, 2\pi\cos\theta_y/\lambda, 2\pi\cos\theta_z/\lambda)\,. \tag{6.41}$$

This enables us to write (6.40) in the rather condensed form

$$\psi(\underline{r}) = A\cos(\underline{k}\cdot\underline{r})\,. \tag{6.42}$$

The fact that $\underline{k}$ is actually a 4-vector follows directly from the fact that its scalar product with $\underline{r}$, which we already know to be a 4-vector, is frame invariant. This is already quite a powerful result since we know the transformation law for 4-vector components, so (6.41) allows us to immediately deduce the transformation laws for ν and λ. However the true utility of this approach is that it enables us to write down coordinate independent, geometric formulae to describe the Doppler effect. The frequency of a wave with wave 4-vector $\underline{k}$, according to an observer with 4-velocity $\underline{u}_{\mathrm{obs}}$ is given by

$$\nu_{\mathrm{obs}} = \frac{\underline{k} \cdot \underline{u}_{\mathrm{obs}}}{2\pi} \,. \tag{6.43}$$

We can see this by noting that it is trivially true in the observer's rest frame, where their 4-velocity is equal to $(c, 0, 0, 0)$, and then using the fact that the scalar product between two 4-vectors is frame invariant to infer that it must hold true in all frames. Thus, if a wave is emitted from some source with 4-velocity $\underline{u}_{\mathrm{em}}$ and received by a detector with 4-velocity $\underline{u}_{\mathrm{rec}}$, the emitted and received frequencies will be related by a ratio of

$$\frac{\nu_{\mathrm{rec}}}{\nu_{\mathrm{em}}} = \frac{\underline{k} \cdot \underline{u}_{\mathrm{rec}}}{\underline{k} \cdot \underline{u}_{\mathrm{em}}} \,. \tag{6.44}$$

This expression can then be evaluated in whatever frame we find most convenient, and upon doing so will immediately give us the same formulae that we derived for the Doppler effect earlier.

Appendix

General Lorentz transformations

In special relativity we generally regard space and time as being homogeneous, meaning that no one event is privileged over any other. It is for this reason that the coordinates t', x', y', z' belonging to an inertial frame $\mathcal{K}'$ must be at most linear functions of the coordinates t, x, y, z belonging to a frame $\mathcal{K}$. Translating the origin of a frame is trivial as it amounts to simply shifting the coordinates by some constant offset. As such, we usually consider two frames which share an origin, in which case the relationship between their coordinates is known as a Lorentz transformation. This Lorentz transformation must be linear, and as such must be expressible in the general form

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \Lambda_{tt} & \Lambda_{tx} & \Lambda_{ty} & \Lambda_{tz} \\ \Lambda_{xt} & \Lambda_{xx} & \Lambda_{xy} & \Lambda_{xz} \\ \Lambda_{yt} & \Lambda_{yx} & \Lambda_{yy} & \Lambda_{yz} \\ \Lambda_{zt} & \Lambda_{zx} & \Lambda_{zy} & \Lambda_{zz} \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \Lambda_{tt}ct + \Lambda_{tx}x + \Lambda_{ty}y + \Lambda_{tz}z \\ \Lambda_{xt}ct + \Lambda_{xx}x + \Lambda_{xy}y + \Lambda_{xz}z \\ \Lambda_{yt}ct + \Lambda_{yx}x + \Lambda_{yy}y + \Lambda_{yz}z \\ \Lambda_{zt}ct + \Lambda_{zx}x + \Lambda_{zy}y + \Lambda_{zz}z \end{pmatrix} . \tag{A.1}$$

Having 4×4 matrices flying around can be a bit hard to follow, so it will be convenient for us to represent these matrices using 1+3 block notation, where we group the three spatial coordinates together. In this notation, (A.1) can be written as

$$\begin{pmatrix} ct' \\ \boldsymbol{r}' \end{pmatrix} = \begin{pmatrix} \Lambda_{tt} & \Lambda_{t\boldsymbol{r}} \\ \Lambda_{\boldsymbol{r}t} & \Lambda_{\boldsymbol{rr}} \end{pmatrix} \begin{pmatrix} ct \\ \boldsymbol{r} \end{pmatrix} = \begin{pmatrix} \Lambda_{tt}ct + \Lambda_{t\boldsymbol{r}} \cdot \boldsymbol{r} \\ \Lambda_{\boldsymbol{r}t}ct + \Lambda_{\boldsymbol{rr}} \cdot \boldsymbol{r} \end{pmatrix} . \tag{A.2}$$

Here $\boldsymbol{r}$, $\boldsymbol{r'}$, $\Lambda_{t\boldsymbol{r}}$, and $\Lambda_{\boldsymbol{r}t}$ are to be interpreted as three component vectors, while $\Lambda_{\boldsymbol{rr}}$ is a 3×3 matrix. The dot products in $\Lambda_{t\boldsymbol{r}} \cdot \boldsymbol{r}$ and $\Lambda_{\boldsymbol{rr}} \cdot \boldsymbol{r}$ refer to the regular vector dot product and matrix multiplication respectively. In this notation the interval can be written as

$$s^2 = \begin{pmatrix} ct & \boldsymbol{r} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -\boldsymbol{1} \end{pmatrix} \begin{pmatrix} ct \\ \boldsymbol{r} \end{pmatrix} = \begin{pmatrix} ct' & \boldsymbol{r}' \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -\boldsymbol{1} \end{pmatrix} \begin{pmatrix} ct' \\ \boldsymbol{r}' \end{pmatrix} . \tag{A.3}$$

Substituting in (A.2) for the primed coordinates, we will find that the Lorentz transformation can only leave the interval unchanged for arbitrary values of t and $\boldsymbol{r}$, if it satisfies the condition that

$$\begin{pmatrix} \Lambda_{tt} & \Lambda_{\boldsymbol{r}t} \\ \Lambda_{t\boldsymbol{r}} & \Lambda_{\boldsymbol{rr}}^T \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -\boldsymbol{1} \end{pmatrix} \begin{pmatrix} \Lambda_{tt} & \Lambda_{t\boldsymbol{r}} \\ \Lambda_{\boldsymbol{r}t} & \Lambda_{\boldsymbol{rr}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -\boldsymbol{1} \end{pmatrix} , \tag{A.4}$$

where a superscript T indicates matrix transposition. Expanding out the matrix multiplication, we obtain the constraint

$$\begin{pmatrix} \Lambda_{tt}^2 - \Lambda_{\boldsymbol{r}t} \cdot \Lambda_{\boldsymbol{r}t} & \Lambda_{tt}\Lambda_{t\boldsymbol{r}} - \Lambda_{\boldsymbol{rr}}^T \cdot \Lambda_{\boldsymbol{r}t} \\ \Lambda_{tt}\Lambda_{t\boldsymbol{r}} - \Lambda_{\boldsymbol{rr}}^T \cdot \Lambda_{\boldsymbol{r}t} & \Lambda_{t\boldsymbol{r}} \otimes \Lambda_{t\boldsymbol{r}} - \Lambda_{\boldsymbol{rr}}^T \cdot \Lambda_{\boldsymbol{rr}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -\boldsymbol{1} \end{pmatrix} . \tag{A.5}$$

The tensor product in this expression implies that the two vectors are combined two create a single 3×3 matrix according to

$$\Lambda_{t\boldsymbol{r}} \otimes \Lambda_{t\boldsymbol{r}} = \begin{pmatrix} \Lambda_{tx} \\ \Lambda_{ty} \\ \Lambda_{tz} \end{pmatrix} \begin{pmatrix} \Lambda_{tx} & \Lambda_{ty} & \Lambda_{tz} \end{pmatrix} = \begin{pmatrix} \Lambda_{tx}\Lambda_{tx} & \Lambda_{tx}\Lambda_{ty} & \Lambda_{tx}\Lambda_{tz} \\ \Lambda_{ty}\Lambda tx & \Lambda_{ty}\Lambda_{ty} & \Lambda_{ty}\Lambda_{tz} \\ \Lambda_{tz}\Lambda_{tx} & \Lambda_{tz}\Lambda_{ty} & \Lambda_{tz}\Lambda_{tz} \end{pmatrix} . \tag{A.6}$$

One thing to note is that, although the 4×4 matrix equation (A.5) has sixteen components, six of those are repeated because both the left and right hand sides are symmetric matrices. As such, there are only really ten true constraints on the sixteen free parameters in (A.1). Thus, we should find that the general formula for a Lorentz transformation has six free parameters which we can choose the values of.

We can begin our analysis of the constraints imposed by (A.5) by invoking a mathematical result known as the polar decomposition theorem. This states that we can always write the matrix $\Lambda_{\boldsymbol{rr}}$ in the form

$$\Lambda_{\boldsymbol{rr}} = \mathrm{R} \cdot \sqrt{\Lambda^T_{\boldsymbol{rr}} \cdot \Lambda_{\boldsymbol{rr}}}\,, \tag{A.7}$$

where R is a 3×3 rotation matrix. As a rotation matrix R must satisfy the conditions that

$$\mathrm{R}^T = \mathrm{R}^{-1} \quad \text{and} \quad \det \mathrm{R} = 1\,. \tag{A.8}$$

There are a number of different parametrisations for rotation matrices; however, we shall provide one it terms of three angles θ, ϕ, ψ. The matrix

$$\mathrm{R} = \begin{pmatrix} c_\theta c_\phi & -s_\phi & s_\theta c_\phi \\ c_\theta s_\phi & c_\phi & s_\theta s_\phi \\ -s_\theta & 0 & c_\theta \end{pmatrix} \begin{pmatrix} c_\psi & -s_\psi & 0 \\ s_\psi & c_\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_\theta c_\phi & c_\theta s_\phi & -s_\theta \\ -s_\phi & c_\phi & 0 \\ s_\theta c_\phi & s_\theta s_\phi & c_\theta \end{pmatrix}, \tag{A.9}$$

where $c_\theta = \cos\theta$ etc., represents a rotation, through an angle ψ, anticlockwise about an axis with polar angles θ and ϕ. This leaves three final degrees of freedom remaining in our parametrisation. Given any three dimensional vector $\Lambda_{t\boldsymbol{r}}$, it will always be possible to find a vector $\boldsymbol{v} = (v_x, v_y, v_z)$ such that

$$\Lambda_{t\boldsymbol{r}} = -\frac{\gamma_v \boldsymbol{v}}{c} = -\frac{\boldsymbol{v}}{\sqrt{c^2 - v^2}}\,. \tag{A.10}$$

This notation is certainly highly suggestive, and indeed we will eventually come to identify $\boldsymbol{v}$ as the relative velocity between the two frames. However, for now we should just view $\boldsymbol{v}$ as a convenient parametrisation for $\Lambda_{t\boldsymbol{r}}$.

Substituting this into (A.5) and looking at the components in the lower right 3×3 block we can see that we must have

$$\Lambda_{t\boldsymbol{r}} \otimes \Lambda_{t\boldsymbol{r}} - \Lambda^T_{\boldsymbol{rr}} \cdot \Lambda_{\boldsymbol{rr}} = -\mathbf{1} \implies \Lambda^T_{\boldsymbol{rr}} \cdot \Lambda_{\boldsymbol{rr}} = \mathbf{1} + \frac{\boldsymbol{v} \otimes \boldsymbol{v}}{c^2 - v^2}\,. \tag{A.11}$$

We now wish to find the square root of this matrix so we can insert it back into our expression for $\Lambda_{\boldsymbol{rr}}$. The easiest way to do this is to simply guess the answer and then check that it works. A sensible looking guess would be something of the form

$$\sqrt{\Lambda^T_{\boldsymbol{rr}} \cdot \Lambda_{\boldsymbol{rr}}} = \mathbf{1} + \zeta \boldsymbol{v} \otimes \boldsymbol{v}\,, \tag{A.12}$$

for some ζ. Squaring this expression yields

$$(\mathbf{1} + \zeta \boldsymbol{v} \otimes \boldsymbol{v})^2 = \mathbf{1} + (2\zeta + \zeta^2 v^2)\boldsymbol{v} \otimes \boldsymbol{v}\,, \tag{A.13}$$

which then implies that

$$2\zeta + \zeta^2 v^2 = \frac{1}{c^2 - v^2} \implies \zeta = \frac{\sqrt{1 + v^2/(c^2 - v^2)} - 1}{v^2} = \frac{\gamma_v - 1}{v^2}\,, \tag{A.14}$$

where we have chosen the positive root of the quadratic because a matrix square root must be positive semi-definite. Thus, we obtain

$$\sqrt{\Lambda_{\boldsymbol{rr}}^T \cdot \Lambda_{\boldsymbol{rr}}} = \mathbf{1} + \frac{(\gamma_v - 1)\boldsymbol{v} \otimes \boldsymbol{v}}{v^2}\,, \tag{A.15}$$

and hence conclude that

$$\Lambda_{\boldsymbol{rr}} = \mathrm{R} \cdot \left(\mathbf{1} + \frac{(\gamma_v - 1)\boldsymbol{v} \otimes \boldsymbol{v}}{v^2}\right)\,. \tag{A.16}$$

We can now examine the off diagonal components of (A.5), which tell us that

$$\Lambda_{tt}\Lambda_{t\boldsymbol{r}} - \Lambda_{\boldsymbol{rr}}^T \cdot \Lambda_{\boldsymbol{r}t} = 0\,. \tag{A.17}$$

Substituting in our known expressions for $\Lambda_{t\boldsymbol{r}}$ and $\Lambda_{\boldsymbol{rr}}$ we find that this can be written in the form

$$-\Lambda_{tt}\frac{\gamma_v \boldsymbol{v}}{c} - \left(\mathbf{1} + \frac{(\gamma_v - 1)\boldsymbol{v} \otimes \boldsymbol{v}}{v^2}\right) \cdot \mathrm{R}^T \cdot \Lambda_{\boldsymbol{r}t} = 0\,. \tag{A.18}$$

Noting that $\mathrm{R}^T = \mathrm{R}^{-1}$ and that

$$\left(\mathbf{1} + \frac{(\gamma_v - 1)\boldsymbol{v} \otimes \boldsymbol{v}}{v^2}\right)^{-1} \cdot \boldsymbol{v} = \frac{\boldsymbol{v}}{\gamma_v}\,, \tag{A.19}$$

we arrive at the conclusion that

$$\Lambda_{\boldsymbol{r}t} = -\frac{\Lambda_{tt}\boldsymbol{v}}{c}\,. \tag{A.20}$$

To find the value of Λ_{tt} we substitute this result into the time component of (A.5), which yields

$$\Lambda_{tt}^2 - \Lambda_{\boldsymbol{r}t} \cdot \Lambda_{\boldsymbol{r}t} = \Lambda_{tt}^2\left(1 - \frac{v^2}{c^2}\right) = 1 \implies \Lambda_{tt} = \frac{1}{\sqrt{1 - v^2/c^2}} = \gamma_v\,. \tag{A.21}$$

Putting this all together we can write the most general Lorentz transformation as

$$\begin{pmatrix} ct' \\ \boldsymbol{r}' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \mathrm{R} \end{pmatrix} \begin{pmatrix} \gamma_v & -\gamma_v \boldsymbol{v}/c \\ -\gamma_v \boldsymbol{v}/c & \mathbf{1} + (\gamma_v - 1)\boldsymbol{v} \otimes \boldsymbol{v}/v^2 \end{pmatrix} \begin{pmatrix} ct \\ \boldsymbol{r} \end{pmatrix}\,. \tag{A.22}$$

The definition $\Lambda_{t\boldsymbol{r}} = -\gamma_v \boldsymbol{v}/c$ is justified by noting that $\boldsymbol{r}' = 0$ if and only if $\boldsymbol{r} = \boldsymbol{v}t$, and so $\boldsymbol{v}$ must be the velocity of the primed frame relative to its unprimed counterpart.

We can rewrite these transformations slightly by introducing a new rotation matrix Q, such that $\mathrm{Q}\boldsymbol{v}$ lies along the x axis. If we also define a second rotation matrix $\mathrm{Q}' = \mathrm{R}\mathrm{Q}^{-1}$, then the transformation takes the form

$$\begin{pmatrix} ct' \\ \boldsymbol{r}' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \mathrm{Q}' \end{pmatrix} \begin{pmatrix} \gamma_v & -\gamma_v v/c & 0 & 0 \\ -\gamma_v v/c & \gamma_v & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \mathrm{Q} \end{pmatrix} \begin{pmatrix} ct \\ \boldsymbol{r} \end{pmatrix}\,, \tag{A.23}$$

demonstrating that any Lorentz transformation can be obtained via the composition of rotations and a boost in the standard configuration.