First Year Special Relativity – Lecture 1
Introduction, concepts and classical results

Mitesh Patel, 12th May 2022

1 In this lecture

- Concepts of active and passive transformations of coordinates;
- Physical laws in different coordinate systems;
- Galilean (pre-Relativity) transformations.

2 Introduction

For the material in this lecture, the Galilean transformations are covered in Young and Freedman, Sec. 3.5 and in McCall in Sec. 5.2.

Special Relativity, as everyone knows, is a theory introduced by Albert Einstein. The original publication was in German; the title can be translated as "Electrodynamics of moving bodies". This title indicates that the theory is closely tied to electromagnetism. When it was published in 1905, the theory was originally just called Relativity and was so named because it describes how physical objects behave when moving *relative* to an observer at fixed velocities. Hence, it does not really cover objects which accelerate, so handling these situations can be tricky. Indeed, eleven years after his original theory of Relativity, in 1916 Einstein came up with a more general theory which also included acceleration. The latter is now called General Relativity and is basically the fundamental theory of gravity, while the original theory has been renamed as Special Relativity, as it is a special case, *i.e.* when gravitational fields (and all other forces which cause accelerations) can be neglected, so all objects have fixed velocities.

We will only study Special Relativity in this course. The mathematics is actually quite easy and contains nothing beyond GCSE-level; there is no calculus, for example. The difficulty is understanding the mind-bending concepts. In contrast, General Relativity is mathematically very advanced, which is why it is a fourth year option course and well beyond these lectures.

3 Rotations

As we will see in later lectures, Relativity involves 'transformations' of coordinates. Rotations are another specific case of transformations; they move positions in space from one point to another. Rotations should be familiar to you and they have a lot in common with the relativistic transformations we will meet later. Hence we can introduce some of the concepts involved in Relativity first using rotations. To simplify the mathematics, we will only consider rotations around the z axis, which therefore mix x and y into each other. Also, for simplicity, we will also only consider rotations around the origin $x = y = 0$.

An 'active' rotation is where we consider an object to be rotated in a fixed coordinate system, often called a coordinate 'frame'. This type of rotation corresponds physically to moving the object. For any given point on the object, like a corner, the x and y values change to new values x' and y' according to

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

where the rotation is by angle ϕ defined relative to the x axis as in standard spherical or cylindrical coordinates. Note that because we rotate around the origin, the point $x = y = 0$

transforms to $x' = y' = 0$ for any rotation angle. It is intuitively clear that to rotate back to the starting point, we need to rotate by the same angle but in the opposite direction. To prove this mathematically, we need to invert the 2×2 matrix in the standard way. Firstly, the determinant Δ is given by

$$\Delta = (\cos\phi) \times (\cos\phi) - (\sin\phi) \times (-\sin\phi) = \cos^2\phi + \sin^2\phi = 1$$

and, since $\cos(-\phi) = \cos(\phi)$ while $\sin(-\phi) = -\sin(\phi)$, the matrix equation can be inverted by multiplying on both sides by the inverse matrix to give

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{\Delta} \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(-\phi) & -\sin(-\phi) \\ \sin(-\phi) & \cos(-\phi) \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}$$

The inverse matrix is seen to be a rotation matrix but with an angle $-\phi$.

There is also a 'passive' rotation, where the coordinate frame is rotated, rather than the object itself. This means the object is completely unchanged. Because the coordinate frame changes, then although the vector position of a corner is unchanged, it appears to have different values for its x and y components in the rotated coordinate frame. It is again intuitively clear that a passive rotation of the coordinate frame by ϕ changes the coordinates x and y in the same way as an active rotation of an object by $-\phi$. Hence, a passive rotation is given by

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Fig. 1 compares the two types of rotation.

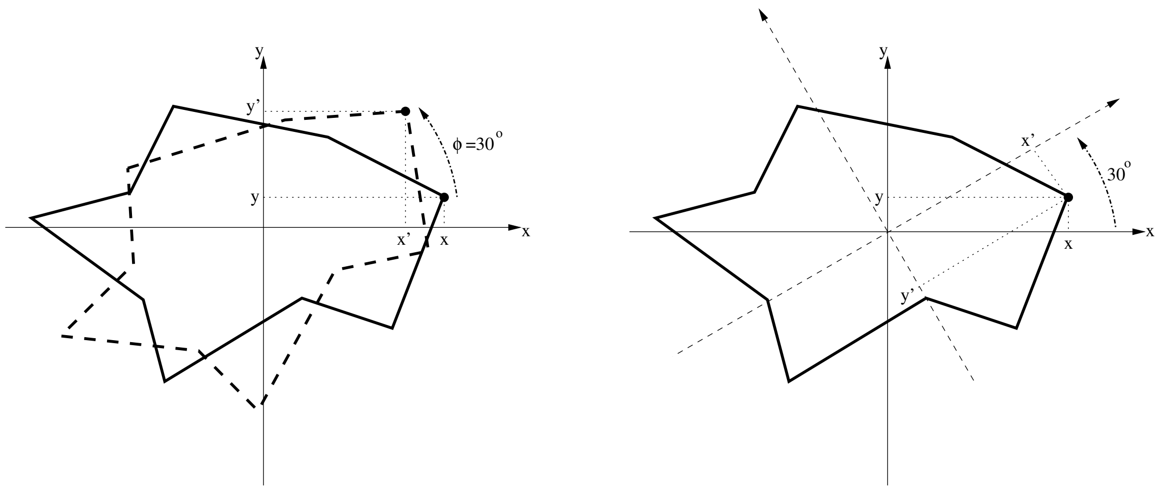

Figure 1: Left: Active rotation by 30° of an object in a fixed coordinate system. Right: Passive rotation by 30° of the coordinate system around a fixed object.

It is a fundamental assumption of physics that space is 'isotropic', which means it is the same in all orientations and has no 'special' direction. Another way to say this is that we believe there is a perfect symmetry in all directions of space.

Consider someone, who we will generally call an 'observer', does an experiment. Afterwards they do an active rotation of the equipment and repeat the experiment. As long as all objects which influence the experiment are included in the move, then the isotropic symmetry of space means that the observer should get the same result the second time. This means all physical laws must still be valid after a rotation. Note, we have one observer and do the experiment twice.

We can consider the same experiment equipment but instead do a passive rotation. To make it effectively identical to the previous case, the passive rotation can be chosen to be the negative angle of the active rotation. Here we do the experiment once, but we have two observers, one in each of the two coordinate frames, with one using the original coordinate frame and another using the rotated frame. It is clear that they will see the same outcome and agree on the result, and hence will deduce the same laws of physics from the experiment. (However, this again only works because the physical laws don't depend on a specific coordinate, e.g. x.) Therefore, when doing these comparisons, we can think in terms of either an active or a passive rotation and as long as there is a symmetry, it doesn't matter which. In Relativity, we will usually think in terms of passive transformations and multiple observers.

There is one more critical concept which arises from the symmetry of space. If physical laws are valid in all frames, then which coordinate frame to use is arbitrary. Any frame is as good as any other frame; no coordinate frame is 'better' than the others. Hence, it is meaningless to think of a 'true' frame which defines $\phi = 0$ because there is no way to define such a frame. Hence, it is impossible for a physics experiment to be able to tell its 'true' ϕ orientation because physical laws work the same for any rotation. Only relative changes in ϕ can have any effect.

4 Invariant and covariant

Let's think about what 'agreement' of the rotated experiment results means in more detail. If the two observers only measure quantities which do not change under rotations, such as mass, temperature, voltage, etc., then it is clear they will write down exactly the same measurements and hence will agree. However, not all physical quantities have this property. They may also measure position, velocity, electric field, etc., which change direction under a rotation and so the two observers will get different values for the three coordinates of each of these quantities.

Given that their values are not identical, how do they check the physical laws are still the same? Clearly, all the quantities in the first group are scalars and the second are vectors. Another word for a scalar is an 'invariant' and this is used more often in Relativity textbooks. For non-native English speakers, in the word 'invariant', the starting 'in' means 'not' and 'variant' (like 'vary') means changable, so an 'invariant' is a quantity that is 'not changable'. Here it means it does not change under rotations. If we only had invariants in physics, then knowing we have the same physical laws for both observers would be trivial. However, it is more complicated; one important example is Newton's second law $\vec{F} = m\vec{a}$, which involves a scalar m but also two vectors $\vec{F}$ and $\vec{a}$. The critical point is that this equation will work in all observer coordinate frames because the two vectors rotate in the same way, *i.e.* both sides of the equation change consistently. Equations like Newton's second law are called 'covariant', where 'co' here means 'together' (e.g. 'cooperate' means 'work together' and 'cohabit' means 'live together'). Hence 'covariant' means both sides of the equation 'change together', *i.e.* in the same way, and hence if Newton's second law holds in one frame, it will hold in all rotation frames. Generalising this, then since we believe all rotation frames are equivalent, all physical laws must be able to be expressed in covariant equations, which have scalars or vectors (or more complicated structures) matching on both sides. Basically, using vectors correctly does this job for us. Fundamentally, scalars and vectors are really *defined* by their transformations under rotations. You already know not to form equations which have different types of quantities on each side, e.g. $\vec{F} = m|\vec{a}|$ must be wrong as it has a vector on one side and a scalar on the other, so it is not covariant.

Note, there are various operations we can do with vectors, such as dot products $\vec{a}.\vec{b}$, including of a vector with itself $\vec{a}.\vec{a} = |\vec{a}|^2$, which give invariants under rotations. Since the vector length (squared) is unchanged when rotated, then by definition, the curve along which each point moves under an active rotation will be the curve with a constant value of length, *i.e.* a circle.

5 Inertial frames and Galilean transformations

There is a fundamental concept underlying Relativity which is equivalent to the rotational symmetry; all physical laws must be valid when transforming from one coordinate frame to another. However, in Relativity, the transformation is not due to a change of angle, but to a change of velocity. Although it is not initially very obvious, the relativistic transformation of coordinates for moving frames is mathematically similar (although clearly not identical) to the mathematics for transformations between rotated frames given above. In rotations, for simplicity we only considered rotations around the z axis. Similarly, when considering inertial frames, we will only consider velocity changes along the x axis.

It is important to remember that Relativity is the special case of frames moving only with fixed velocities. Such coordinate frames are called 'inertial frames' because Newton's first law (which still holds even with Relativity) says the inertia of an object means it will continue to be at rest, or move with a fixed velocity, if no force is applied to it. An inertial coordinate frame is one in which this statement holds so the frame itself must also be at rest or moving with a constant velocity.

We should first see how this works in classical physics, *i.e.* before Relativity was discovered. How do objects appear to observers in different inertial frames? Classically, it seems obvious that if an object has a speed u along the $+x$ axis in one inertial frame, then in a different inertial frame moving with a constant speed v along the $+x$ axis, it will appear to have a speed

$$u' = u - v$$

The position in x of the object in the first frame is related to u by $u = dx/dt$ and similarly in the second frame $u' = dx'/dt$. Hence, integrating the relationship above (and remembering that v is a constant) gives

$$x' = x - vt + C$$

for some constant of integration C. In the same way as we always rotated around the origin, in this course we will always do transformations which make the origins of the two coordinate systems agree at $t = 0$. Hence to get $x' = x = 0$ at $t = 0$ clearly means we take $C = 0$. To make the transformations more similar to what we will see later, they can be written as

$$t' = t, \qquad x' = x - vt, \qquad y' = y, \qquad z' = z$$

These are called the passive Galilean transformations, named after Galileo. (The active transformations simply change the sign of v.) Before the 19^{th} Century, all known physical laws held in all moving frames if the Galilean transformations were used. In particular, Newton's laws are covariant under the Galilean transformations; for instance, energy and momentum will have different values in different inertial frames but in any frame, an observer will measure their values will be conserved.

The Galilean transformations can be contrasted with the passive rotation transformations written out in a similar format

$$t' = t, \qquad x' = x\cos\phi + y\sin\phi, \qquad y' = y\cos\phi - x\sin\phi, \qquad z' = z$$

Critically, note the Galilean transformations do not look particularly like a rotation between x and t because there is no x term in t'. We will discuss the modified transformations used in Relativity in Lecture 4.

The Galilean transformations are intuitive as they correspond to our everyday experience. However, it turns out they are not completely correct. Instead, they are only an approximation to the exact relativistic transformation laws. Many of the conceptual problems in Relativity arise from the fact that our intuition expects the Galilean transformations and so it can initially be hard to understand the consequences of the exact relativistic transformations.

6 Revision questions

- What is the difference between an active and passive rotation?
- In cases where the physical laws don't depend on a specific direction, what do we expect for the results of some experiment after such rotations?
- Is the rotated or original coordinate frame more correct to make calculations in?
- In general, will an observers interpretation of events depend on which frame they observe an experiment from?
- What is a covariant equation?
- What is an inertial frame?
- What are the Galilean transforms and how will energy and momentum transform between inertial frames?

First Year Special Relativity – Lecture 2
The postulates of Relativity

Mitesh Patel, 16th May 2023

1 In this lecture

- Look at the inconsistency between Maxwell's equations and the Galilean transformations;
- The Michelson-Morley experiment – no evidence for a medium in which light propagates to help reconcile the above inconsistency;
- The postulates of Relativity;
- First consequence of these postulates – time dilation.

2 Introduction

The material in this lecture is covered in Young and Freedman, Secs. 37.1 and 37.3 and in McCall in Secs. 5.3, 5.6 and 5.7.

Having seen the example of rotational coordinate transforms and what classical physics assumed about relative motion, we will now look at why the Theory of Relativity arose and its postulates.

3 Maxwell's equations and the aether

One of the great successes of the discovery of what are now called the Maxwell equations was that they showed that light is waves of electric and magnetic fields. In particular, the equations in vacuum can be used to give a wave equation

$$\epsilon_0\mu_0\frac{\partial^2\vec{E}}{\partial t^2} - \nabla^2\vec{E} = 0$$

In a wave equation, the leading constant is $1/c^2$, where c is the wave speed. This showed that the speed of light was directly related to two other quantities, the vacuum permittivity ϵ_0 and the vacuum permeability μ_0, as measured in electrostatic and magnetic experiments (respectively) in the laboratory, by $c^2 = 1/\epsilon_0\mu_0$, which fixes c. Hence, if Maxwell's equations are correct, c is a fundamental constant of nature and all light must travel at this speed in vacuum.

However, as opposed to all previously known physics laws, it turns out Maxwell's equations are not covariant under the classical Galilean transformation equations discussed in the previous lecture. The easiest way to see this is that the Galilean transformations change speed from u to $u' = u - v$, so this would imply that the speed of light should depend on the speed of the observer v. To reconcile these two apparently conflicting results, the idea of a medium called the 'aether' (sometimes spelt 'ether') was raised. This was based on an analogy with sound; this also has a speed in a medium (for sound, this is the air) calculable from the properties of the medium. However, sound does appear to have different speeds to moving observers. This is because the speed is relative to the medium and so sound only goes at the calculated speed in the frame where the medium is at rest. In other inertial frames, its speed will be dependent on its direction. Hence, the medium picks out a special inertial frame and breaks any inertial frame symmetry. Consequently physical laws will appear to be different in different frames and we would be able to tell if we were in the special frame where the medium is not moving (called 'at rest') as only then would the speed of sound would be the same in all directions.

The aether was postulated to be the medium for light waves; it was the substance that was oscillating as the waves passed through the medium. As we can see light from distant stars, the aether was assumed to fill the whole Universe, even in the vacuum of outer space where there is nothing in terms of normal matter. Hence, as for sound, we would see light travels at a different speed than $1/\sqrt{\epsilon_0\mu_0}$ if we were moving relative to the frame where the aether is stationary and its speed would depend on direction. Effectively, the aether theory assumes the Galilean velocity transformation laws are correct but that the equation for the speed of light, and hence also Maxwell's equations, only hold in one frame but are otherwise incorrect. This is of course testable; the Earth moves in orbit round the Sun with a speed of around 30 km/s. Measurements of the relative speed of light in two perpendicular directions taken at several different angles should show a difference in light speed when one direction aligns with the relative motion of the Earth and the aether. The difference is small ($\sim 10^{-4}$ of c) compared to the speed of light but measurable using a two-arm interferometer. The first such experiment sensitive enough to see a change in speed of this magnitude was done in 1887 by Michelson and Morley. They found no such effect and hence concluded that the aether theory must be wrong. This is the classic case of a 'failed' experiment which turned out to be extremely important, and really shows how critical is it to eliminate hypotheses as well as make discoveries.

With the aether theory rejected, Einstein realised the resolution of this dilemma was the other way round. He proposed that Maxwell's equations were correct, and hence $c^2 = 1/\epsilon_0\mu_0$ is always true for all observers, but it was the Gailiean transformation equations which were wrong. Specifically, they needed to be changed so that light is observed as going at c in every inertial frame and the way Einstein found to do this was by allowing the *time* to be different in different frames. By changing time in just the right way, then the speed of light can appear to be the same for all frames.

4 Postulates of Relativity

Any theory must have some postulates (i.e. assumptions) from which the rest of the results arise. The theory then stands or falls on whether these assumptions are correct and this can only be verified experimentally. Relativity has only two such postulates:

1. The laws of physics are the same in all inertial frames, i.e. coordinate frames moving uniformly relative to each other.

2. The speed of light c in vacuum is independent of the speed of the light source and has the same value for all inertial observers.

Although Einstein published the Theory of Relativity over a hundred years ago, it is still the case that no violation of these two principles has ever been measured experimentally.

The first postulate is often itself called the 'Relativity Principle' as it concerns relative motion and says you can't tell how fast you are going in absolute terms. This is the reason we discussed rotations in the first lecture. In both cases, we consider how any system (such as a physics experiment) looks to different observers who are using different coordinate systems. Different inertial frames are analogous to different rotated frames and hence this is a statement that all laws of physics are the same ('covariant') under transformations between the coordinate systems of different inertial frames. It means there is a fundamental symmetry between all inertial frames. Also, as for rotations, the first postulate means it is impossible to build a physics experiment which can detect if it is moving in absolute terms; only relative motion is meaningful. It should be emphasised again that an inertial frame must have a constant velocity; acceleration itself can be easily detected, even directly by your body with no equipment needed.

The second postulate is basically a statement that Maxwell's equations are correct. As discussed above, the Michelson-Morley experiment found no effect due to an aether which would have indicated that Maxwell's equations were only approximate. We now believe the Maxwell's equations are exact (until quantum mechanics is included) and indeed they are covariant under Einstein's relativistic transformations. This was first noticed by Lorentz and this is why the actual transformations we will discuss in Lecture 4 are named after him. However, Lorentz treated the covariance of Maxwell's equations as being due to the specific properties of electromagnetic forces and still assumed that the Galilean transformations were correct. It was left to Einstein to fully understand their implications more generally.

We should be careful of the term 'observers' as used in the second postulate. For rotations of coordinate frames, there is no big conceptual problem here. We would assume the two people observing an experiment using two different coordinate systems would be able to see the whole experiment at all times. However, in Relativity, there is often some initial confusion about the word 'observer' as it does imply visually seeing something, and Relativity is all about the speed of light. However, this is incorrect; when we say 'observer' we actually mean the person has some system that allows them to record everything happening in all space at all times. It is as if they have video cameras everywhere throughout space recording time-stamped images and the observer can look at these recordings at their leisure. Adding in the delay due to the finite speed of light to work out what they would actually visually see from one particular point in space in real time is a whole extra layer of complications. The visual appearance of relativistic objects to a single person is in fact quite an interesting study but not what we will consider in this course.

We will assume we can go to any inertial frame and there will be an (imaginary) observer who is at rest in the coordinate system of that frame. Hence, we can 'observe' any system from any speed we like, just as we could observe a system from any angle by using passive rotations. In the jargon, changing from one inertial frame to another is often referred to as doing a 'boost'. This implies an acceleration, which would be needed if the system was physically speeded up, i.e. by an active transformation, but in Relativity we will usually be considering passive transformations.

What happens to the origin in one frame relative to the other? By definition, if an inertial frame is moving at speed v relative to another frame, then the origin (or any other point) in the moving frame has speed v and so will move a distance vt in time t. Hence, an object at rest in that inertial frame will also have speed v. From the perspective of the observer in this frame, the original frame is moving in the opposite direction, but also with the same speed. The speeds could not possibly be different because, as we saw in the previous lecture, space is isotropic and there is nothing which could tell us which direction would be higher and which lower.

5 The light clock

We can get a very important result on the differences between time in different inertial frames very simply by considering a 'light clock'. This is an example of a 'thought experiment', where the consequences of a particular system are thought through, as performing a real experiment would be very difficult. A light clock is conceptually simple device which uses light bouncing off a mirror to measure time. A pulse of light is emitted, reflected by the mirror a distance d away, and then detected when it returns. The time taken is then the clock period T. Clearly, in the clock rest frame, the period is $T = 2d/c$.

Consider another frame where the clock moving at velocity v perpendicular to the separation of the light emitter and mirror. Figure 1 shows the situation in the rest frame and the moving frame.

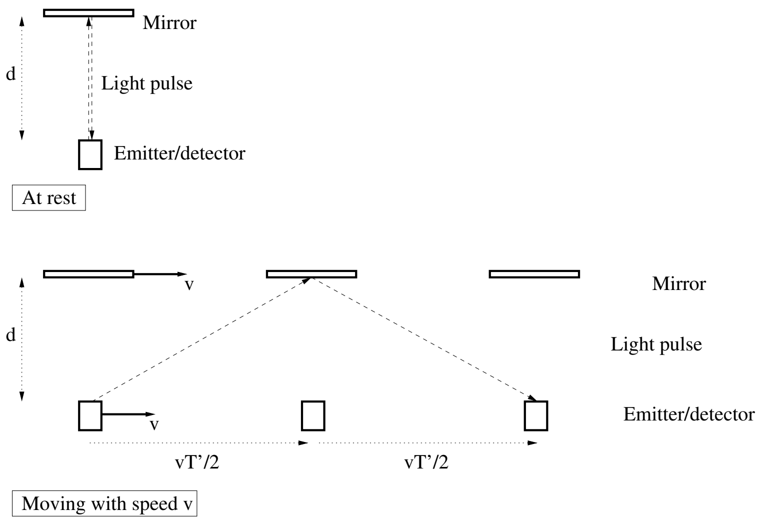

Figure 1: The lightclock at rest (top) and moving to the right with speed v (bottom).

The light pulse clearly has to go further in the moving frame. However, from the postulates of Relativity, it travels at speed c in all frames so the time taken must be longer; let the moving frame period be T'. By symmetry, the time taken for the light to get to the mirror is $T'/2$. In this time, the clock moves $vT'/2$ while the light goes $cT'/2$ and simply by Pythagoras

$$\left(\frac{cT'}{2}\right)^2 = \left(\frac{vT'}{2}\right)^2 + d^2$$

The solution is

$$T' = \frac{2d}{\sqrt{c^2 - v^2}} = \frac{T}{\sqrt{1 - v^2/c^2}} = \frac{T}{\sqrt{1 - \beta^2}} = \gamma T$$

where we have defined two new dimensionless symbols which will be used throughout the course

$$\beta = \frac{v}{c}, \qquad \gamma = \frac{1}{\sqrt{1 - \beta^2}}$$

These are both dimensionless; β is the velocity of the clock as a fraction of c (or in other words, in units of c), while γ can be considered to be a function of β. Another combination which often appears is $\gamma\beta$ and a useful relation to remember is

$$\gamma^2 - \gamma^2\beta^2 = \gamma^2(1 - \beta^2) = 1$$

which is obvious from the definition of γ above. Figure 2 shows how γ and $\gamma\beta$ depend on β.

Note $\gamma = 1$ for $\beta = 0$ and increases with $|\beta|$, going to $\gamma \to \infty$ as $\beta \to \pm 1$. If $|\beta| > 1$, then γ would become imaginary and hence so would T'. There is no physical meaning to this result so we will assume from now on that nothing can go faster than c. This applies both to the clock (if we were doing an active transformation) and to the observer (if we were doing a passive transformation). We will return to this point several times during the course.

6 Time dilation and proper time

Going back to the result on the clock periods, as $\gamma > 1$ for any moving frame, then $T' > T$, as stated earlier. This means that the clock has a longer period, i.e. is running slower, by a factor

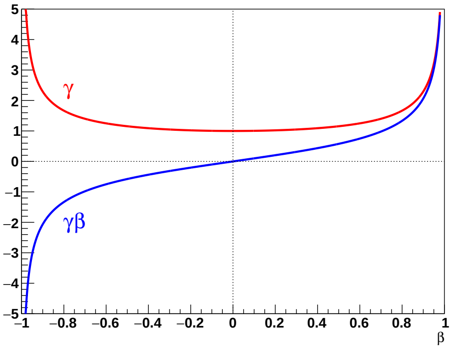

Figure 2: The functions γ and $\gamma\beta$ as a function of β.

of γ in any moving frame compared with the rest frame of the clock. One issue is that as all inertial frames are equivalent, the same is true the other way around! Any observer in the rest frame of the clock will see time in the moving frame running slower than in the rest frame by the same factor. This seems completely contradictory; the solution arises from the fact that time does not only depend on the relative speed of the inertial frames but also on the object position in space, as we will see when we introduce the Lorentz transformations in Lecture 4.

This effect is called 'time dilation', as time has slowed down for the moving object. You might think this is an artifact of the light clock itself, as it uses light which is clearly has a special role in Relativity. However, imagine a different type of clock, e.g. an old-fashioned alarm clock, also at rest in the light clock rest frame and right next to it. These will obviously remain in synchronisation and will tick at the same time in the rest frame. The pair of clocks as seen in a moving frame must give the same result, i.e. must tick synchronously, as it is the same system being viewed by different observers. This means the clockwork mechanism must also run slower by a factor of γ. Clearly, this must be true for every possible time-measuring device which anyone can ever dream up and this must include biological aging. Hence, we conclude the effect of time dilation is not related to the physical structure of the clock or any other object but is intrinsic to time itself. Time really does slow down for objects moving relative to an observer in another frame.

The minimum period of the light clock is clearly when $\gamma = 1$ for which $T' = T$. This is when $\beta = 0$ so the observer is in the rest frame of the clock. The time in the rest frame is therefore the shortest between two ticks. This time is called the 'proper time' (although there is no implication in the name that the time in other frames is in any sense 'improper', meaning not correct or not real) and the usual symbol for this is τ. Since any object, such as a human being, considers itself to be in its own rest frame, then the proper time measures the time as experienced by the object.

7 Revision questions

- Why was the aether was proposed and how did we conclude it did not exist?
- What are the postulates of Relativity?
- What is the smallest possible value of the relativistic gamma parameter?
- How is the period, T', of a clock observed in a moving frame S' related to it's period, T, in it's rest frame, S? Does this effect depend on the type of clock?

- How do we reconcile this with the fact that an observer in the frame S' would observe her own clock to be running slower (longer period) than that in the frame S, which from her perspective would be the one moving *i.e.* she would observe $T' < T$?
- What is the proper time?

First Year Special Relativity – Lecture 3
Length contraction and simultaneity

Mitesh Patel, 18th May 2023

1 In this lecture

- Length contraction;
- Non simultaneity.

2 Introduction

The material in this lecture is covered in Young and Freedman, Secs. 37.2 and 37.4 and in McCall in Secs. 5.9 and 5.10.

We have seen that time runs slower for an object when it is moving relative to an external observer, the so-called 'time-dilation' effect. This lecture discusses some related effects. The first is called 'Lorentz contraction' (or 'length contraction'), in which an object also get shorter along its direction of motion. The second is that occurrences which happen at the same time (i.e. are simultaneous) for one observer are not necessarily simultaneous for other observers.

3 Lorentz contraction

We considered the light clock in Lecture 2, for which the period T is lengthened to $T' = \gamma T$ when it is moving. In fact, *any* clock will have exactly the same dilation factor. We can now ask how far does the clock move between each tick, i.e. within the time of the period.

Consider two inertial frames; one where the clock is at rest and another where the clock is moving with speed v. An observer in the moving frame of the clock has a long wooden rod and cuts this so that, as the clock moves past it, the rod is just the right length so the clock is at one end when it first ticks and reaches the other end when it ticks a second time; see Fig. 1. Note, this observer is in the rest frame of the wooden rod. Since the clock has speed v and the period in the moving frame is T' the observer will find the stationary wooden rod has a length $l = vT'$.

Now consider how this appears to an observer in the clock rest frame. This observer sees the clock as stationary but sees the wooden rod moving with speed v (but in the opposite direction, of course). The clock ticks must still occur just as the start and end of the rod pass the clock as it is the same system being observed. For this observer, the length of the moving wooden rod must be $l' = vT = vT'/\gamma$ and so

$$l' = \frac{l}{\gamma}$$

This is the Lorentz contraction equation. Note, the rod is stationary in the first frame considered, i.e. the moving frame of the clock, not the frame where the clock is at rest. Hence, the length in the rest frame of the rod is l, sometimes called the 'proper length', and in a moving frame is $l' = l/\gamma$, which is shorter. We conclude that lengths get shorter when an object is moving. This is called 'length contraction', or 'Lorentz contraction', as the objects appear to shrink compared to their length at rest. Although not shown by this argument, it is only the length in the direction of motion which is shortened.

You might think that this happens purely due to the particular molecular forces of this particular rod. However, the argument above is very general so it does apply to all objects. In

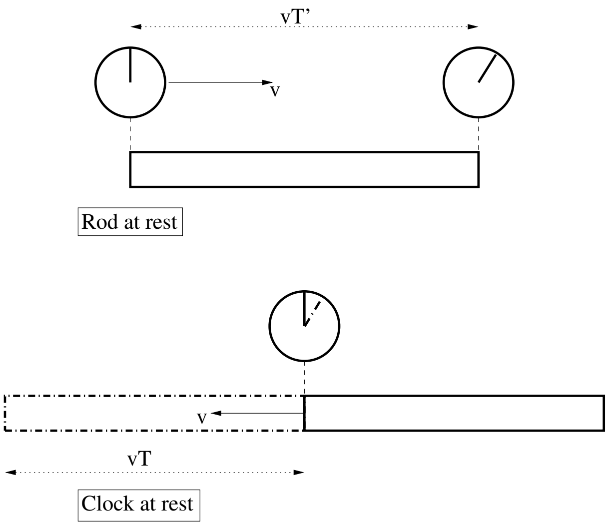

Figure 1: Top: A wooden rod cut to be the same length as the distance the clock moves in one period, shown in the rest frame of the rod. Bottom: The same system in the rest frame of the clock.

fact, it is even more general. Instead of trimming a rod to the right length, we could instead hold flags at the positions where the clock ticks. We would see the same contraction of the distance between the flags but in this case, there is no physical object between the flags. Hence, we have to conclude that it is space itself, and hence any object or distance in that space, which has contracted. This sometimes leads to confusion; what is the 'rest frame' for a distance if there is no object? The distance is defined by its beginning and end (marked by the flags in the case above) and by definition these have to have the same velocity or the distance would be changing with time. The 'rest frame' for the distance is then the inertial frame in which the beginning and end positions of the distance are not moving.

To summarise: time dilation means time slows down for an object in a moving frame compared with its rest frame, while Lorentz contraction means lengths get shorter for an object in a moving frame compared with its rest frame.

As should be clear from the above derivation, length contraction and time dilation are of course inextricably linked. They are two aspects of the same effect and a system when viewed in different frames can use one or the other to explain what is happening.

4 The light clock revisited

Now we know about Lorentz contraction, we can now look at the light clock again but with a different orientation. We previously studied it in a frame moving perpendicular to the direction of motion of the light. Now consider it in a frame moving parallel to the motion of the light, as shown in Fig. 2.

In the clock rest frame, the total time for the light pulse to return is $T = 2d/c$. The distance between the light source and the mirror will be shortened to d/γ but the mirror is moving at speed v. Hence, the time for the light to reach the mirror t'_1 is given by

$$ct'_1 = \frac{d}{\gamma} + vt'_1, \qquad \text{so} \qquad t'_1 = \frac{d}{\gamma(c-v)} = \frac{d}{\gamma(1-\beta)c}$$

Since the source and detector are also moving, the time t'_2 for the light to return from the mirror

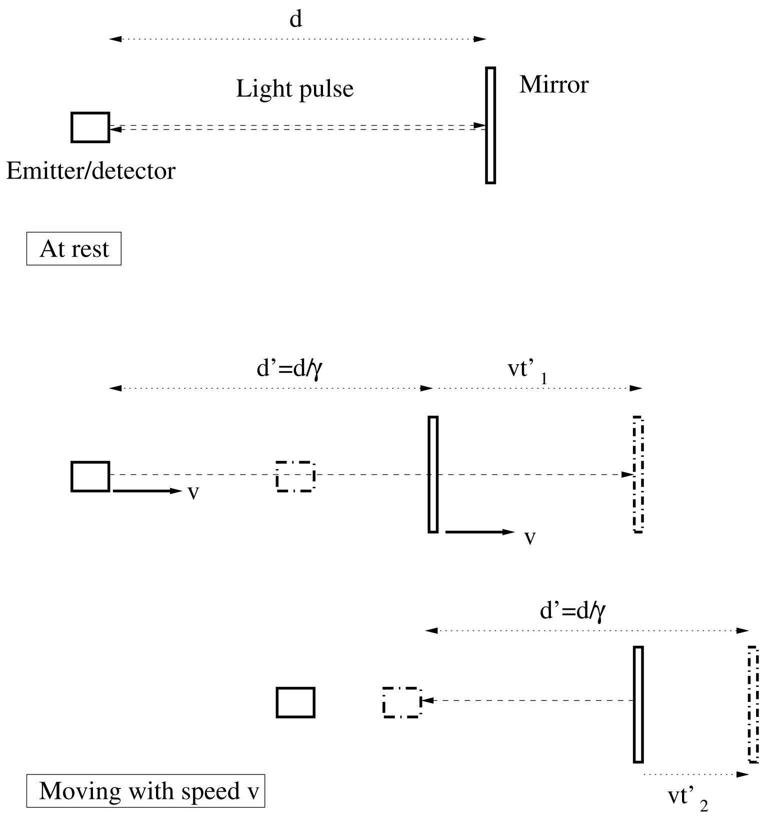

Figure 2: The lightclock at rest (top) and moving to the right with speed v (bottom), oriented so that the light pulse is along the direction of motion.

to the source is

$$ct'_2 = \frac{d}{\gamma} - vt'_2, \qquad \text{so} \qquad t'_2 = \frac{d}{\gamma(c+v)} = \frac{d}{\gamma(1+\beta)c}$$

Hence, the total time for the light to return, i.e. the clock period, is

$$T' = t'_1 + t'_2 = \frac{d}{\gamma(1-\beta)c} + \frac{d}{\gamma(1+\beta)c} = \frac{d(1+\beta)+d(1-\beta)}{\gamma(1-\beta)(1+\beta)c} = \frac{2d}{\gamma(1-\beta^2)c} = \frac{2d\gamma}{c} = \gamma T$$

Hence, the time dilation factor does not depend on which way round the clock is oriented, as must be the case for consistency; time must not depend on whether we do a rotation of the clock in its rest frame.

However, one interesting feature is that in the perpendicular frame, the time from the source to the mirror and from the mirror to the detector were both equal to $d\gamma/c$, but in this orientation, the two times t'_1 and t'_2 are no longer equal. You might have thought each leg of the light path should simply be time dilated by γ but this is not the case. This is related to the fact that the two ends of the light path are (obviously) not in the same position, i.e. that the observed time depends on position as well. We will see how this works in general in Lecture 4.

5 Simultaneity

Consider a slightly more complicated light clock in its rest frame, where an emitter sends light in two opposite directions to bounce off two mirrors and return to be detected. If the two

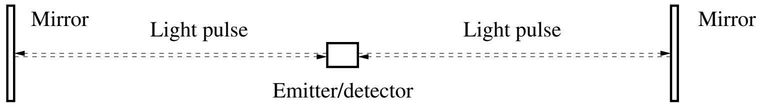

Figure 3: A double light clock, with both arms having an identical length.

mirrors are the same distance from the source then the two light pulses will arrive at the mirrors simultaneously and similarly return to the detector at the same time; see Fig. 3.

How does this look in a moving frame? As the light has speed c, then with one mirror moving towards and the other away from the source, then the light will no longer hit the two mirrors at the same time. By effectively an identical calculation to the above, you can show the times to reach the front and back mirrors are

$$\frac{d}{\gamma(1-\beta)c} \qquad \text{and} \qquad \frac{d}{\gamma(1+\beta)c}$$

respectively. However, having bounced off the mirrors, the light then reaches the detector after

$$\frac{d}{\gamma(1+\beta)c} \qquad \text{and} \qquad \frac{d}{\gamma(1-\beta)c}$$

respectively, which clearly sum to the same total so the light pulse does arrive back at the detector at the same time.

The critical thing here is that the arrival times of the light pulses at the mirrors are simultaneous in the rest frame of the clock, but are not simultaneous in any other inertial frame. However, the arrival at the detector is simultaneous in all inertial frames. The difference is that the light arriving at the mirrors is at two different locations, while the arrival at the detector is at the same location. The general principle is that for two simultaneous occurrences which are not at the same x position, then in any other frame moving along x, they are no longer simultaneous. In contrast, if they *are* simultaneous and at the same x position, they will be simultaneous in all frames moving along x. Hence, to be simultaneous in any frame moving in any direction, the two occurrences must be at exactly the same position.

The breaking of simultaneity raises significant issues. In particular, can the first occurrence affect the second, given that they may occur in the opposite order in some frames to others? We will discuss this in Lecture 6.

6 Revision questions

- Why do we get length contraction?
- Are two simultaneous occurrences that occur at different x positions in some frame also simultaneous in a frame moving along x?
- What about two simultaneous occurrences at the same x position?
- How does the relativity of simultaniety help understand e.g. length contraction?

First Year Special Relativity – Lecture 4
The Lorentz transformations

Mitesh Patel, 19th May 2023

1 In this lecture

- The concept of an event;
- The Lorentz transformations;
- The velocity transformations.

2 Introduction

The material in this lecture is covered in Young and Freedman, Sec. 37.5 and in McCall in Sec. 5.11 and 5.12.

We have seen that the postulates of Relativity lead to several unexpected consequences, like time slowing down, objects contracting and things not being simultaneous for different observers. We now want to look at the general equations which transform points in space and time as all the other effects actually result from the application of these general transformations. These equations are called the 'Lorentz transformations'.

3 Events

When we looked at rotations, the equations change a point at x, y to x', y', i.e. they move a specific point in space. We will see that the Lorentz transformations are mathematically similar and they must also work on a particular point. However, this is not a point just in space but in space and time. Like rotations, Lorentz transformations can also be active or passive, where active would require physically changing the speed of an object (in the jargon, doing a 'boost'), while passive is considering how the object would appear to observers moving at different speeds. If just considering observers moving along the x axis, then we actually only have to consider values of t and x. Hence, we need to consider something which happened at a specific x at a particular time t. Such an occurrance is called an 'event'. This is a fundamental concept in Relativity and as long as you can break anything down into a list of events, you can usually solve any Relativity problem. Note, a true event occurs (in principle) at an infinitessimallly small point in space at an instant in time. Obviously, as long as events are small and short compared with the scales of the system being considered, then they are a good approximation to the ideal.

In terms of events, we assume the 'all-seeing' observer is able to measure the t and x of any event which occurs. Also, we can now make more precise statements about proper time and length. Proper time is the time between two events in an inertial frame in which the events have the same position. Similarly, proper length is the distance between two events in the frame in which the events have the same time and in which the object they mark the ends of is stationary.

4 The Lorentz transformations

Given an event at t, x, y, z, then a passive Lorentz transformation from an initial inertial frame to the frame of an observer moving at speed v along the $+x$ axis can be written as

$$t' = \gamma\left(t - \frac{vx}{c^2}\right), \qquad x' = \gamma(x - vt), \qquad y' = y, \qquad z' = z$$

where $\gamma = 1/\sqrt{1-(v/c)^2}$ as defined previously. The values of t' and x' tell you the coordinates of the event in the different frame. One critical thing to note is that the change from t to t' does not only depend on the speed (through v and so also γ) but also on the position x. This is the basic cause of many of the tricky effects in Relativity, such as non-simultaneity.

Since y and z are unchanged by an observer moving along x, we will usually not bother to write them except when required. Note, in a similar way to rotating around the origin, for simplicity we have chosen the origin so that $t = 0, x = 0$ always transforms to $t' = 0, x' = 0$. Using the definition introduced previously of $\beta = v/c$, so $\gamma = 1/\sqrt{1-\beta^2}$, the transformations can also be written as

$$t' = \gamma\left(t - \frac{\beta x}{c}\right), \qquad x' = \gamma(x - \beta ct)$$

It is obvious that t and x have different dimensions and this makes the equations a little less intuitive. Obviously, as distance is speed times time, then ct has the same dimensions as x. You will have come across this concept before; e.g. 'light-years' is an explicit ct unit. Using ct, we can rewrite the above by multiplying the first equation throughout by c

$$ct' = \gamma(ct - \beta x), \qquad x' = \gamma(x - \beta ct)$$

which then shows the symmetry between ct and x quite clearly. We can also write these two equations as one matrix equation

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma\beta \\ -\gamma\beta & \gamma \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix}$$

so that the similarity to a rotation is more obvious. Clearly this is *not* actually a rotation though. All the elements of a rotation matrix are sines and cosines and so must have a magnitude of one or less. However, we know γ has a minimum (not maximum) of one and can be arbitrarily large. This means $\gamma\beta$ can also be very large too. Also note that the off-diagonal terms have an equal magnitude, but in a rotation matrix they have opposite signs, whereas here they have the same sign. Figure 1 compares the way points under rotations and Lorentz transformation change.

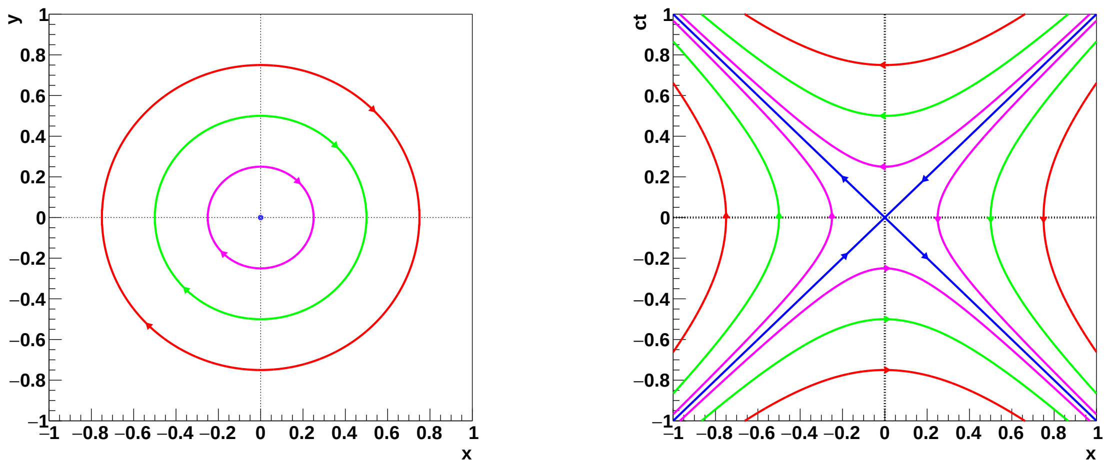

Figure 1: Left: Change of x, y positions under rotations. Right: Change of ct, x events under Lorentz transformations. In both cases, a position or event on one of the lines will move along that line under the relevant transformation.

How would we do the inverse transform to change back to the original observer? As for the rotation case we can invert the matrix. Using the usual 2×2 method, we first find the determinant

$$\Delta = (\gamma)(\gamma) - (\gamma\beta)(\gamma\beta) = \gamma^2 - \gamma^2\beta^2 = \gamma^2(1-\beta^2) = 1$$

where the last step is simply due to the definition of γ. Hence

$$\begin{pmatrix} \gamma & -\gamma\beta \\ -\gamma\beta & \gamma \end{pmatrix}^{-1} = \frac{1}{\Delta}\begin{pmatrix} \gamma & \gamma\beta \\ \gamma\beta & \gamma \end{pmatrix} = \begin{pmatrix} \gamma & \gamma\beta \\ \gamma\beta & \gamma \end{pmatrix}$$

Since $\gamma(-\beta) = 1/\sqrt{1-(-\beta)^2} = 1/\sqrt{1-\beta^2} = \gamma(\beta)$ then, just as for rotations, the inverse is simply a transformation with the same magnitude but opposite sign, i.e. $\beta \to -\beta$, which is what intuition would tell us.

5 Small speed approximation

We also need to be sure an Lorentz transformation corresponds to the classical Galilean transforms for small boosts. If β is small enough, then $\gamma \approx 1$, which means a Lorentz transformation is approximately

$$ct' \approx ct - \beta x, \qquad x' \approx x - \beta ct$$

Since $\beta = v/c$, the second equation can be written as

$$x' \approx x - vt$$

which is what we got for x' classically. The first equation can be written as

$$t' \approx t - \frac{v}{c}\frac{x}{c}$$

and so has an extra term compared with the classical expression $t' = t$. However, for the small speeds we are used to, $v \ll c$, and for our normal day-to-day experiences, any distance x gives an extremely small time value when divided by c. Hence, this extra term consists of the product of two very small numbers and so $t' \approx t$ to a very good approximation, as assumed in the Galilean transformations.

6 Consistency of the speed of light

We are assuming the Lorentz transformation is the correct relativistic transformation. If this is true then it must obey the second postulate, specifically the speed of light must be the same after the transformation. Say (for convenience) a light pulse is generated from the origin $x = 0$ at $t = 0$ and travels along the x axis so after a time T, it will be at $x = cT$. Hence, take its position at $t = 0$ as one event and at $t = T$ as a second event. Since the coordinates are zero for the first event, this event in another frame also has the coordinates at zero. For the second event, then

$$cT' = \gamma(cT - \beta cT), \qquad x' = \gamma(cT - \beta cT)$$

The speed in the boosted frame is $u' = \Delta x'/\Delta T' = x'/T'$ since the first event is at zero, so

$$u' = \frac{x'}{T'} = \frac{\gamma(cT - \beta cT)}{\gamma(cT - \beta cT)/c} = c$$

Therefore, in the boosted frame, the two events are precisely the right distance apart for the light pulse to travel between them in time T'. Hence an Lorentz transformation preserves the speed of light and so does not violate the second postulate.

7 Velocity transformation

It is important to understand that an Lorentz transformation by $\beta = v/c$ does *not* mean all velocities change from u to $u' = u - v$ as in the Galilean transformations. An explicit case is the speed of the light pulse, found above, which does not change speed at all.

Let's do the similar calculation to the light speed one above, but for an object moving at a speed $u < c$. In time T it goes from the origin to uT. Hence, applying an Lorentz transformation to this second event gives

$$cT' = \gamma(cT - \beta uT), \qquad x' = \gamma(uT - \beta cT)$$

The speed in the boosted frame is $u' = x'/T'$ which is

$$u' = \frac{x'}{T'} = \frac{\gamma(uT - \beta cT)}{\gamma(cT - \beta uT)/c} = \frac{u - \beta c}{1 - \beta u/c} = \frac{u - v}{1 - uv/c^2}$$

This is the velocity transformation formula. Note, although calculated from the Lorentz transformations, it is *not* itself a Lorentz transformation; velocity does not change in the same way as position.

A few other points should be noted:

1. For a small velocity $u \ll c$ and small boost $v \ll c$, then $uv/c^2 = (u/c)(v/c)$ is extremely small and hence $u' \approx u - v$, which again agrees with the Galilean result.

2. If the initial velocity is $u = 0$, then $u' = (-v)/1 = -v$ i.e. boosting to a frame going at v along the $+x$ axis makes an object initially at rest appear to have $-v$, as we would expect. Similarly, a boost by $-v$ makes an object initially at rest appear to have velocity $+v$. Also, if the boost velocity is chosen so $v = u$ then $u' = 0$, i.e. we have boosted to catch up with the object so it then appears at rest.

3. If the original velocity was in fact the speed of light, i.e. $u = c$, then $u' = (c - v)/(1 - cv/c^2) = c(1 - v/c)/(1 - v/c) = c$ and so is unchanged, as we found before.

8 Revision questions

- What is an event in the Relativistic sense?
- What feature of the Lorentz transformations gives rise to non-simultaniety?
- What is the form of the Lorentz transformations?
- How do velocities transform between frames? What happens when $u = 0$ or $u' = 0$?

First Year Special Relativity – Lecture 5
Space-time diagrams and world lines

Mitesh Patel, 22nd May 2023

1 In this lecture

- A way of visualising Relativisitic events - space-time diagrams;
- How to draw the trajectory of objects in such diagrams - world lines;
- Transforming between frames in such diagrams.

2 Introduction

The material in this lecture is not covered in Young and Freedman and in McCall only briefly in Sec. 6.1.

In the previous lecture, we say how the Lorentz transformations change a given position in space and time, called an 'event', from one inertial frame to another. This can be represented graphically in a 'space-time' diagram. The trajectory of an object can be considered as a sequence of events. These events form a 'world line' for the object, which is the graphical representation of the trajectory in a space-time diagram.

3 Space-time diagrams

A space-time diagram is simply a visual illustration of the ct and x values of events. By convention, the axes are drawn with the ct axis vertical and x axis horizontal. An event is a single point in such a diagram. A space-time diagram was already included in Lecture 4, which showed the lines along which events will move when they are Lorentz transformed.

The Lorentz transformations defined in Lecture 4 are for passive transformations, where the axes are changing, not the events themselves. This is equivalent to a passive rotation; all we are doing is changing the coordinate system, not the physical object. We saw under a passive rotation that the x and y axes rotate by the same angle. This is not the same for a Lorentz transformation, but we can find how the ct and x axes change under a Lorentz transformation in a space-time diagram quite easily. Specifically, any event on the x' axis by definition has $t' = 0$. This means all such events must satisfy

$$ct' = 0 = \gamma(ct - \beta x) \qquad \text{so} \qquad ct = \beta x$$

This is a straight line in the space-time diagram with gradient β, i.e. it makes an angle α to the original x axis, where $\tan\alpha = \beta$. Similarly, any point on the ct' axis has $x' = 0$ so now all such events must satisfy

$$x' = 0 = \gamma(x - \beta ct) \qquad \text{so} \qquad ct = \left(\frac{1}{\beta}\right)x$$

which is a straight line with a gradient of $1/\beta$. Since $1/\tan\alpha = \tan(\pi/2 - \alpha)$, this is an angle of $\pi/2 - \alpha$ to the x axis, which corresponds to an angle of α to the original ct axis. Note that the gradient in both cases is positive so this does *not* look like the rotation case and the two axes are not perpendicular to each other. Fig. 1 compares the two types of transformation.

You are probably not familiar with how to resolve a vector in non-perpendicular coordinates into its components. The trick is to always move parallel to the 'other' axis. Hence, you can see

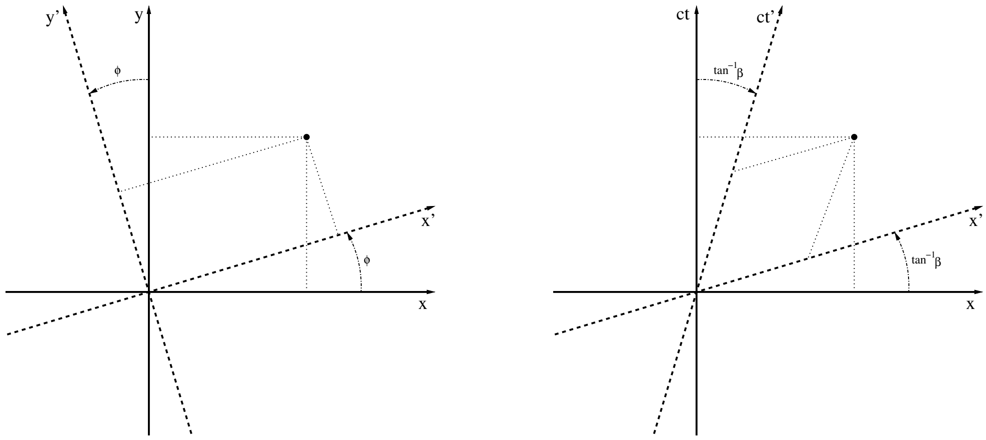

Figure 1: Left: Passive rotation by ϕ of the coordinate system with a given space position. Right: Passive Lorentz transformation by β of the coordinate system with a given space-time event.

in the figure how the new components for the given event correspond to those the second observer will measure. That this is required to resolve components is straightforward to understand. As stated above, any event on the x' axis has $t' = 0$, i.e. the x' axis is a line of constant t', in this case $t' = 0$. Lines of constant t' with values other than 0 will also be parallel to the x' axis but higher or lower, depending on the constant value, as shown in Fig. 2. Any line of constant t' (or t) is referred to as a 'line of simultaneity', because all events on that line are simultaneous in the primed (or unprimed) frame.

In a similar way, lines of constant x' are all parallel to the ct' axis, with the line $x' = 0$ actually corresponding to the ct' axis itself, also shown in Fig. 2. Note that any object at rest in the primed frame will by definition have the same x' value for all times. Hence lines of constant x' correspond to objects at rest in the transformed frame.

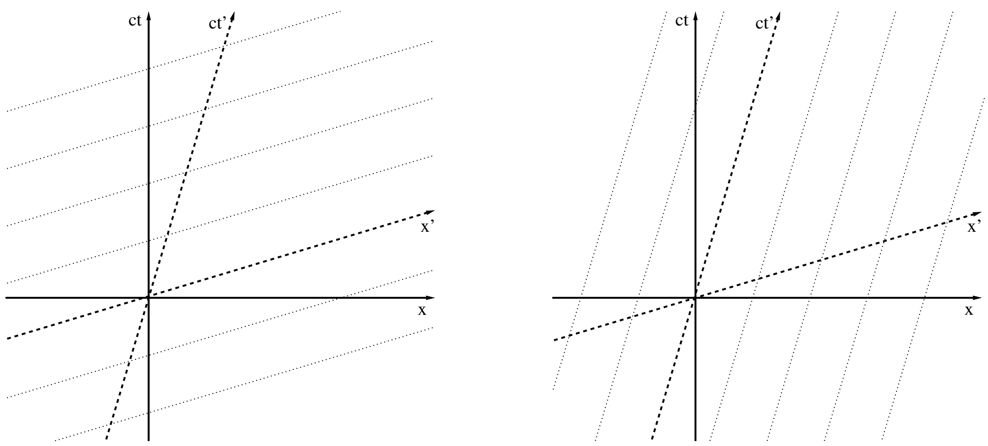

Figure 2: Left: Lines of constant t', also called lines of simultaneity. Right: Lines of constant x', corresponding to world lines of objects at rest in the transformed frame.

Of course, the second observer always considers their axes to be perpendicular and so will view the situation as shown in the right-hand diagram of Fig. 3. From the perspective of the second observer, the original frame coordinates both have negative gradients, of $-\beta$ and $-1/\beta$, as this figure shows. This view is of course equivalent to an inverse Lorentz transformation, just as it was for rotations.

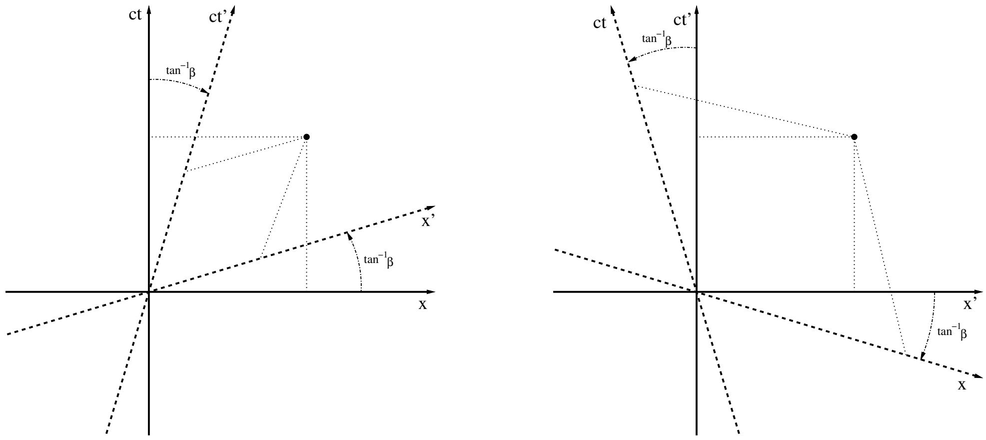

Figure 3: Left: Passive Lorentz transformation by β of the coordinate system with a given space-time event (as previously shown in Fig. 1). Right: View of same transformation from the perspective of an observer in the transformed frame.

4 World lines

You will have solved mechanics problems involving Newton's laws for the position of an object as a function of time, i.e. $x(t)$. It is common to plot the x as a function of t, for example as shown in the left of Fig. 4. For any point on the curve, the derivative of the function gives the velocity, while the second derivative gives the acceleration. Hence, an object without any acceleration has a straight line function (so the second derivative is zero) while a stationary object has a horizontal line (so the first derivative is also zero). The same type of graph is very useful in Relativity also. However, as we saw above, in Relativity space-time diagrams are standardly drawn with the axes swapped over; i.e. the ct axis is vertical and the x axis is horizontal, also shown in Fig. 4.

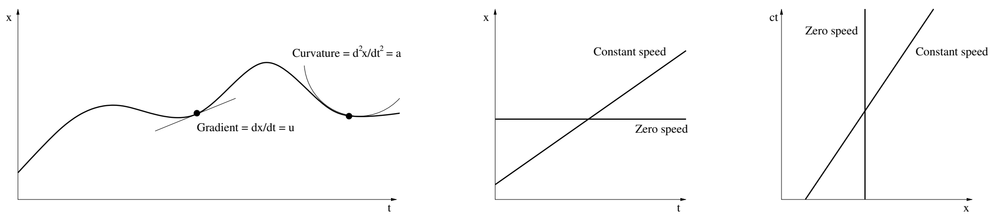

Figure 4: Left: Example of solution $x(t)$ for a mechanics problem, showing velocity and acceleration. Middle: Lines for zero and non-zero constant velocity. Right: Lines for zero and non-zero constant velocity in a space-time diagram, which has the axes exchanged.

Any object which exists for a finite time will form a line in the diagram, just like the one described above; this called the 'world line' of the object. In this case, because the axes are swapped, then a stationary object will give a vertical line while an object moving at constant velocity u will give a straight line with some finite gradient, now given by the inverse of $\beta_u = u/c$.

For the simple case of an object going through the origin, you can think of this explicitly as

$$x = ut = \left(\frac{u}{c}\right) ct = \beta_u ct \qquad \text{so} \qquad ct = \left(\frac{1}{\beta_u}\right) x$$

Since we are limited to speeds $|u| \leq c$ for which $|\beta_u| \leq 1$ and hence $1/|\beta_u| \geq 1$, then the magnitude of the gradient must be always at least 1 and for light itself, it must always travel along lines with a gradient = 1, which have an angle of 45° in space-time diagrams. As stated at the beginning of the course, we will not consider acceleration so we only have to deal with cases like these.

5 Transforming world lines

An object which is stationary in one frame will not be stationary in a different frame but will be seen as moving with a constant velocity. Hence, a vertical world line will become tilted, but still be a straight line, in a different frame. More generally an object moving at a constant speed in one frame has a different constant speed in another frame so again it will remain a straight line but its gradient will have changed in the second frame. You can think of such a line as being a lot of events, each being the position of the object at a different time, with the times all very close to each other. As we can consider a world line as simply a lot of events, then applying the Lorentz transformations to these events will tell us directly how the world lines transform. In practise, since we are only dealing with straight lines, we only need to transform two events, as the new world line will be the straight line going through both of these. Alternatively we can use the velocity transformation formula presented in Lecture 4 to determine the new inverse gradient, although we still need to Lorentz transform one event on the world line to get the new intercept.

The rest frame case is particularly straightforward. An object at rest in a frame has a world line parallel to the ct axis in that frame. Under a Lorentz transformation, in the new frame, the ct axis will be tilted as shown in the right of Fig. 3. Hence, the world line must also have the same gradient as the ct axis in this frame. This is illustrated in Fig. 5.

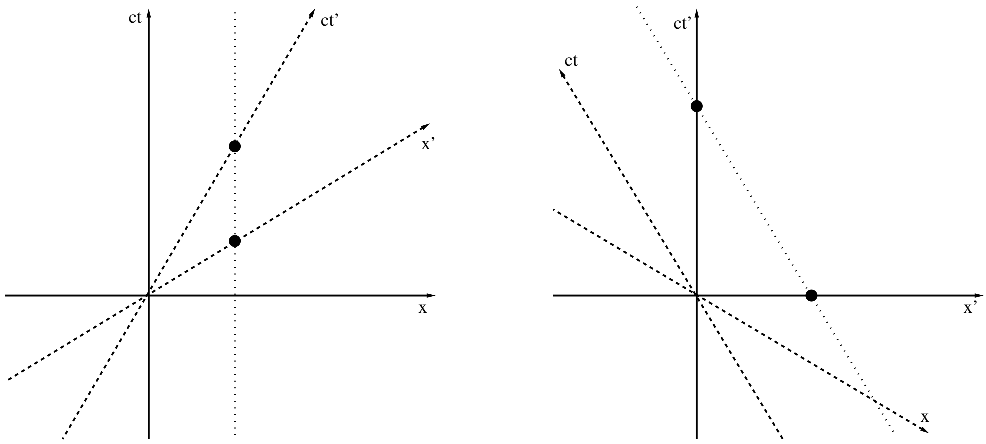

Figure 5: Left: A world line (dotted) for an object at rest in an initial frame. It passes through the ct' and x' axes at the two events marked by circles. Right: The transformed view of the world line, showing the transformed positions of the two events and the transformed world line going through these. Note this world line is still parallel to the original ct axis.

Overall, you should have got the impression that time and space are working in very similar ways within Relativity. In some ways, this is the big conceptual leap; time and space are mixed under Lorentz transformations and the inertial frame equivalence symmetry tells us there is a symmetry between space and time.

A final note. There is a health warning in using the non-perpendicular axes of the transformed frames; the overall γ factor in the Lorentz transformations means the axes are stretched compared with the original axes as well as tilted. This means, unlike for rotations, a distance of e.g. 1 m along the x axis is not the same length in a space-time diagram on the x' axis. Hence you often cannot use simple geometry to compare distances between inertial frames. You should always work in terms of events to be safe.

6 Revision questions

- On a space-time diagram draw:
  - a zero speed object;
  - an object travelling at c;
  - an unphysical (superluminal) line.
- Redraw your diagram from the point-of-view of a moving observer;
- Can you reproduce the world-line diagram to explain why both the tortoise and hare think they won their race?

First Year Special Relativity – Lecture 6
Four-vectors and causality

Mitesh Patel, 23rd May 2023

1 In this lecture

- Four vectors;
- The separation between two events;
- Implications for causality.

2 Introduction

The material in this lecture is not covered in Young and Freedman. and partially in McCall in Secs. 5.8 and 5.13.

We have seen that ct and x mix into each under a Lorentz transformation along the x axis, quite like x and y do under a rotation around the z axis. We know that vectors have well-defined properties when rotated. We can take the analogy of Lorentz transformations and rotations further and combine time and space into a 'four-vector'.

3 Four-vectors

We form the 'space-time' four-vector from ct and $\vec{r}$ and we will write this vector as $(ct, \vec{r})$ or (ct, x, y, z). This is an equivalent notation to writing the normal position vector as (x, y, z). To be absolutely clear which vectors I mean, I will use 'three-vector' for the standard vectors you have met before Relativity. The four-vector $(ct, \vec{r})$ indicates a 'position' in the 4D space-time (sometimes called 'Minkowski space'). Clearly, a specific four-position is what we mean by an event, so another way to think of a Lorentz transformation is that it changes a four-vector in a well-defined way, just like a rotation changes a three-vector. In fact, the full equation for a Lorentz transformation along the x axis in terms of all four components of a four-vector is

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}, \qquad \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\phi & \sin\phi & 0 \\ 0 & -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}$$

while the second equation is the full rotation case, for comparison.

When we refer to the coordinates of a three-vector, we often write $\vec{v} = (v_x, v_y, v_z)$ or (v_1, v_2, v_3). For four-vectors, the time coordinate is written before the space coordinates and to preserve the numbering of the space coordinates, ct gets the label of '0', called the 'zeroth' coordinate. In fact, the notation used by researchers is mainly $(ct, \vec{r}) = (x^0, x^1, x^2, x^3)$ although I will not use this here; it can be a little confusing as x is being used for all four components.

As you know, the length of a three-vector $|\vec{v}|$ does not change under a rotation. Obviously therefore neither does the square of the length $|\vec{v}|^2 = v_x^2 + v_y^2 + v_z^2$. Similarly there is a length-squared for a four-vector but it is not quite what you might initially think. For the space-time four-vector, it is given by $S^2 = (ct)^2 - |\vec{r}|^2 = c^2t^2 - x^2 - y^2 - z^2$. Hence it does not simply sum all four of the squared components but you have to remember to subtract the space coordinate terms from the time coordinate term. Unlike a three-vector, the length-squared of a four-vector can therefore be positive, zero or negative. The reason for the S^2 definition is, as for a three-vector under rotations, the length-squared of a four-vector is unchanged by a Lorentz transformation. Since y and z don't change, the calculation to show this reduces to

$$\begin{aligned}(ct')^2 - x'^2 &= \gamma^2(ct - \beta x)^2 - \gamma^2(x - \beta ct)^2 = \gamma^2(c^2t^2 - 2\beta ctx + \beta^2x^2 - x^2 + 2\beta ctx - \beta^2c^2t^2) \\ &= \gamma^2[c^2t^2(1-\beta^2) - x^2(1-\beta^2)] = (ct)^2 - x^2\end{aligned}$$

As we discussed for rotations, values which do not change are called scalars or invariants and so the length-squared S^2 of a four-vector is an invariant, or more specifically a 'Lorentz invariant'.

We saw the curves along which events move under Lorentz transformations in Lecture 3; this is reproduced below, together with the rotation case for comparison. We now understand that those correspond to lines of constant length-squared, i.e. constant $c^2t^2 - x^2$ (given that y and z don't change). The diagram with these curves labelled by their Lorentz invariant values shows that the different parts of the figure have different signs for the length-squared.

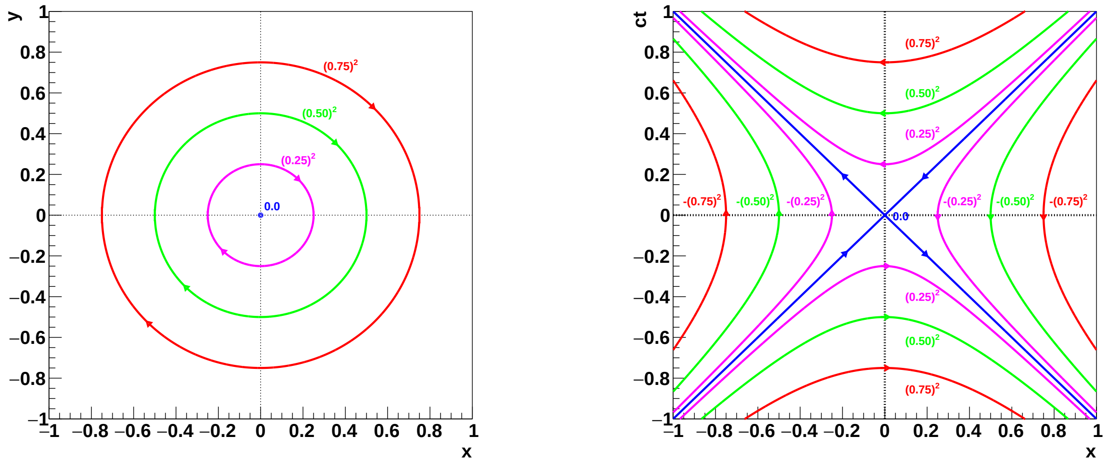

Figure 1: Lines of constant length-squared for rotations (left) and Lorentz transformations (right). The values of the length-squared in both cases are indicated.

Beware: there are several other notations and conventions. The above is closest to the modern notation used within General Relativity and particle physics. However, there are others. One is to call ct the fourth component and then the invariant interval is $x^2+y^2+z^2-c^2t^2$ which is changed in sign compared to the previous definition. There is even a notation where the fourth component is made imaginary and written as ict, so that the negative sign in the invariant interval is 'automatically' taken into account by simply summing the squares. This sounds attractive, but General Relativity generalises the constants $(+1, -1, -1, -1)$ to be variables. Hence, using ict only handles the Special Relativity case.

4 Event separation

Just as for three-vectors, we can do arithmetic with four-vectors. Two four-vectors can be added or subtracted by adding the coordinates separately. If we want the distance between two points $\vec{r}_1$ and $\vec{r}_2$ in three-vector space, then we know this distance squared is $(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2$. We can consider this as taking the vector difference of the two vectors $\Delta\vec{r} = \vec{r}_2 - \vec{r}_1$ and then taking the length-squared $|\Delta\vec{r}|^2$. It should be clear that under rotations, $\Delta\vec{r}$ will behave just like all other three-vectors.

Again, the same holds for four-vectors and events. The 'separation' between two events is defined to be the length-squared of the four-vector resulting from subtracting the four-vectors of the two events. This four-vector is sometimes written as $(c\Delta t, \Delta\vec{r}) = (ct_2 - ct_1, \vec{r}_2 - \vec{r}_1)$. This difference is also a four-vector and so obeys the Lorentz transformations. The separation is $\Delta S^2 = c^2\Delta t^2 - |\Delta\vec{r}|^2$ and is also clearly a Lorentz invariant. If we consider two events, one at the origin $t = 0, \vec{r} = 0$ and the other at ct and $\vec{r}$ then the separation is $\Delta S^2 = c^2t^2 - |\vec{r}|^2$ and so corresponds to what we discussed earlier in the lecture. Events with different signs of the separation have very different properties with respect to each other.

5 Causality

The sign of the invariant interval between two events is crucial for understanding one of the most important concepts in Relativity, namely 'causality'. The base of this word is 'cause' and it concerns whether one event can cause, or more generally affect, another. If they can, they are said to be 'causally connected'.

Consider two events where the second is close in position to the first, but later in time. It is obvious that the first can affect the second. How far apart can the two events be such that the first can actually affect the second? We have assumed nothing can go faster than light so if an object travelling at a speed up to the speed of light from the first event can get to the second event, then clearly the first can affect the second. Hence, we require the distance to be less than or equal to the distance light can go in that time. The events must satisfy $|\Delta\vec{r}| \leq c\Delta t$ which means $|\Delta\vec{r}|^2 \leq c^2\Delta t^2$ or $\Delta S^2 = c^2\Delta t^2 - |\Delta\vec{r}|^2 \geq 0$, i.e. the invariant interval must not be negative. This is the condition for one event to be able cause the other.

Let's turn the argument around: when would two events *not* be able to affect each other? One obvious case is when they are separated in space but happen at exactly the same time. There is clearly no way any object or signal can move between two separated events instantaneously without going faster than c, which we assume is not allowed. The more general case is that a signal even at the speed of light would not get between them in the time available. This means the two events must satisfy $|\Delta\vec{r}|^2 > c^2\Delta t^2$ or $\Delta S^2 = c^2\Delta t^2 - |\Delta\vec{r}|^2 < 0$. Hence, as you might have guessed, the condition is that the invariant interval is negative. These events are called 'causally unconnected'.

One worry is that because events have different times in different inertial frames, could one appear to affect the other in one frame and not the other? The answer is no; since the condition depends on the invariant interval, which is the same in all frames, then the ability of one event to affect the other is the same in all frames. Clearly this must be absolutely required for physical laws to obey any logic. In fact, the speed limit of c applies not just to physical objects but in fact even to information. If the events can communicate in any way using faster-than-light signals, then there will be a frame where an earlier event was caused by a later event in that frame, which breaks all logic. Hence, the requirement of causality is very strong and limits all things which can influence events not to go faster than light.

Let's look at the Lorentz transformation diagram in Fig. 1 again. With one event at the origin, the other will move along the lines of constant separation. The whole of the central upper quadrant of the figure has a positive separation so all events there can be affected by an event at the origin. Note, the time of the second event is always positive, i.e. later than the time of the first event. You can always find a particular boost which moves the second event along the line until it is directly above the first event. This puts the second event at $x = 0$, i.e. so the two events only differ in time and not space. More generally, you can change the position order of the events, so if the first event is to the right of the second in one frame, it can become to the left of the second event in another frame. The events in the lower central quadrant also have a positive separation and here their times are always earlier than the origin; in this case they can cause the event at the origin but not vice versa. Any events which can influence, or be influenced by, a specific event are said to be within the 'light-cone' of that event, illustrated in Fig. 2. The name comes from considering the shape in two space dimensions (as it is hard to visualise in all three).

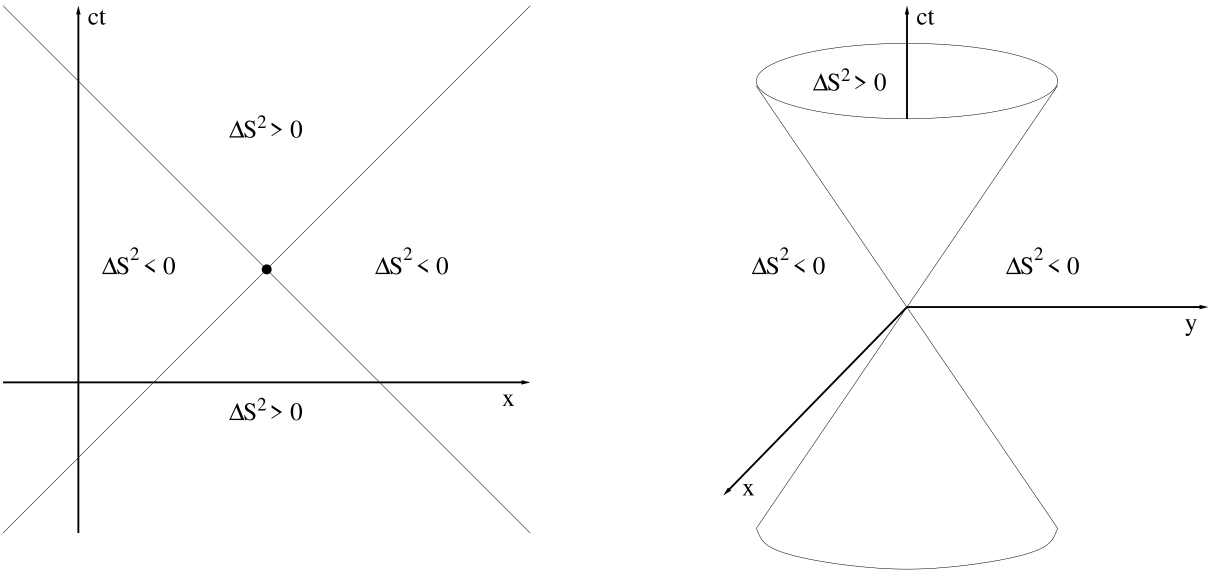

Figure 2: Left: The light-cone for the event marked by a circle and corresponding separations in ct, x. Right: 3D visualisation of light-cone in ct, x, y for an event at the origin.

What about the two quadrants to the left and right in Fig. 1? These have a negative separation from the origin and so the events cannot affect each other. Here, the time difference of the events can change between being positive and negative under boosts. Hence these are the ones for which the time order of the events changes in different frames. Indeed, there is always one particular boost which moves the event along the curve until it makes the time difference zero so that the two events are simultaneous in that frame. However, they will not be simultaneous in any other frame. The one exception is both events being at $x = 0, t = 0$ which are then simultaneous in all frames. Note that events cannot move from the left quadrant to the right quadrant, so the x order is always preserved. In summary:

1. For $\Delta S^2 > 0$, the time order of two events is the same in all frames and the first can affect the second. These events are called 'time-like' separated because there is always one frame in which they have no difference in position, but are separated in time. The separation is therefore $\Delta S^2 = c^2\tau^2$ where τ is the proper time between the events. The space order is different in different frames.

2. For $\Delta S^2 < 0$ the time order of two events is different in different frames and so they cannot logically affect each other. The space order is always preserved. These events are called 'space-like' separated because there is always one frame in which they have no difference in time but are separated in space.

3. There is a final, special case where ΔS^2 is exactly zero. The events are connected only by a light-speed signal and lie exactly on the diagonals of the light cone. They retain their time (and indeed space) order for any boost. In principle the first can affect the second; hence they are causally connected, as for time-like separated events. However, because c is the same in all frames, they always lie on the diagonals and there is no frame in which they

have either a zero time or a zero space difference. These are called 'light-like' separated, for obvious reasons.

Note, the time dilation formula can only be applied to time-like separated events because for space-like separated events, there is no frame in which the events are at the same position. Similarly, the length contraction formula can only be applied to space-like separated events, because time-like separated events cannot happen at the same time.

6 Tachyons

We have said it is essential for logical consistency that if one event causes the other, that this event happens first in every possible inertial frame. This seems to be true from the above so everything is fine.

However, imagine two events where a particle going *faster* that the speed of light moved from the first to the second. Such hypothetical particles have a general name of 'tachyons' (from the Greek 'tachy' meaning 'rapid'). Although we have assumed this doesn't happen, this does not sound so bad; the first event still happens before the second so it does sound logical. However, if the two events are connected by a particle going faster than light then they must have a negative separation. This means the order of the two events is different in some frames than in others. If you define several events along the tachyon world line and Lorentz transform them to such a frame, you will find the tachyon is apparently going in the reverse direction along the world line i.e. backwards in time compared to the original frame. This does lead to logical inconsistencies and so it is generally assumed tachyons cannot exist; they have certainly never been found experimentally.

7 Revision questions

- Write out a space-time four vector in 4 component terms and in abbreviated 2 component form;
- For a generic 4-vector, what quantity is invariant under Lorentz transformations?
- What does the sign of the separation, S, imply between two events?
- Sketch the ct, x plane for lines of constant S, for an event at the origin in this plane, label the different ΔS regions;
- Which regions are "time-like", "space-like" and "light-like"?

First Year Special Relativity – Lecture 7
Energy and momentum

Mitesh Patel, 26th May 2023

1 In this lecture

- Four-vector for E,p and its dependence on an objects speed;
- Check low speed approximations;
- Look at invariant length-squared;
- Look at the special case of zero mass.

2 Introduction

The material in this lecture is covered in Young and Freedman, Secs. 37.7 and 37.8 and in McCall in Secs. 6.1 and 6.2.

There are many three-vectors in physics; position, velocity, momentum, electric and magnetic fields, etc. There would be no point in generalising the concept to four-vectors if the only one which existed was $(ct, \vec{r})$. Hence, as you probably already guessed, there are many other four-vectors. In the same way that all three-vectors have the same rotation transformation, then all four-vectors undergo the same Lorentz transformation under a given boost. We shall look at one of the other most important four-vectors over the next few lectures, namely the combination of energy and momentum. As we have seen, $\vec{r}$ (a three-vector) is combined with t (which is a scalar under rotations) and these mix together under a Lorentz transformation; i.e. t goes with $\vec{r}$. Similarly E (a scalar) goes with $\vec{p}$ (a three-vector).

However, it is a common mistake to think every three-vector must have another variable to go with it; there are many three-vectors which do not. One is velocity; we already saw this does *not* transform using the Lorentz transformations but in a more complicated way, without any other scalar variable being involved. Another example is the electric and magnetic fields; in this case they together form a structure called a 'tensor' and under boosts, the two vectors mix into each other. The most obvious example of this is that a stationary point charge has a electric field but no magnetic field. However, in the frame of an observer moving relative to the stationary charge, the charge appears to be moving and you should know from the E&M course that a moving charge, i.e. a current, has a magnetic field associated with it.

3 Energy and momentum

The energy E and momentum $\vec{p}$ of any object form a four-vector $(E, \vec{p}c)$, called the 'four-momentum'. As for the time component in the $(ct, \vec{r})$ case, the c ensures the dimensions of all the components are the same. (Warning: some books define four-momentum with the c in other positions, e.g. $(E/c, \vec{p})$.) Since all four-vectors change in the same way, then when changing frames p_y and p_z are unchanged, while

$$\begin{pmatrix} E' \\ p'_x c \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma\beta \\ -\gamma\beta & \gamma \end{pmatrix} \begin{pmatrix} E \\ p_x c \end{pmatrix}$$

Clearly in the rest frame $\vec{p} = 0$ but E cannot also be zero or the above equations would give zero energy and momentum in all frames. Let the rest frame energy be E_0. Consider a passive transformation of the object at rest (i.e. $u = 0$) in the x direction by v, so that the object we are considering is moving in the opposite direction with $u' = -v$. We can define $\beta'_u = u'/c = -\beta$ and $\gamma'_u = 1/\sqrt{1-\beta'^2_u} = \gamma$. Dropping the subscript of p_x for clarity, we get

$$E' = \gamma E_0 = \gamma'_u E_0, \qquad p'c = -\gamma\beta E_0 = \gamma'_u \beta'_u E_0$$

We can find the value of E_0 by requiring consistency with the classical momentum. For small β'_u, then $\gamma'_u \approx 1$ so $p' = \gamma'_u \beta'_u E_0/c \approx u' E_0/c^2$. But we know non-relativistically that for a particle moving at speed u in the x direction $p = mu$ so we conclude $E_0/c^2 = m$, i.e.

$$E_0 = mc^2$$

which is an equation you possibly might have seen before. This energy is often called the 'rest energy' or 'rest mass energy'. Its implications are profound and affect all kinematics, as well as resulting in nuclear energy and weapons; this will be discussed in the next few lectures.

Now we have the energy in the rest frame, we can find the exact formulæ for energy and momentum in any other frame. Substituting into the above gives

$$E' = \gamma'_u mc^2, \qquad p'c = \gamma'_u \beta'_u mc^2 \quad \text{so} \quad p' = \gamma'_u m u'$$

Since these equations only involve quantities all measured in the same frame, we will drop the primes from now on. The change in energy when moving, compared to being at rest, is by definition what we mean by kinetic energy K. Hence for a moving object, generally $K = E - mc^2$ which can also be written as $K = (\gamma_u - 1)mc^2$.

For a velocity $\vec{u}$ in any direction, the momentum term above obviously generalises to $\vec{p} = \gamma_u m\vec{u}$. These equations give E and $\vec{p}$ in terms of $\vec{u}$, but it is useful to also have the inverse relations. Since

$$\vec{u} = \frac{\gamma_u mc^2 \vec{u}}{\gamma_u mc^2} = \frac{\vec{p}c^2}{E} \qquad \text{then} \qquad \vec{\beta}_u = \frac{\vec{u}}{c} = \frac{\vec{p}c}{E}$$

It is also sometimes convenient to have γ_u and $\gamma_u \vec{\beta}_u$ which are

$$\gamma_u = \frac{\gamma_u mc^2}{mc^2} = \frac{E}{mc^2} \qquad \gamma_u \vec{\beta}_u = \frac{\vec{p}c}{E}\frac{E}{mc^2} = \frac{\vec{p}c}{mc^2}$$

One comment on the expression for $E = \gamma_u mc^2$. As $u \to c$, then $\gamma_u \to \infty$ and hence $E \to \infty$ too. Therefore, this illustrates yet another reason why we consider nothing can go faster than light; it takes infinite energy to speed an object up to light speed, so achieving $u = c$ is not possible, let alone $u > c$.

4 Small speed approximation

We already used $p \approx mu$ for small speeds, but we can also ask what is the energy in this approximation? We need to improve on simply $\gamma_u \approx 1$ or else we will just get $E \approx mc^2$ again. We can get a better approximation for γ_u by considering it as a binomial expansion, for which

$$(1+a)^b \approx 1 + ab$$

Since $\gamma_u = (1-\beta_u^2)^{-1/2}$, then with $a = -\beta_u^2$ and $b = -1/2$

$$\gamma_u = (1-\beta_u^2)^{-1/2} \approx 1 + \left(-\frac{1}{2}\right)\left(-\beta_u^2\right) \approx 1 + \frac{1}{2}\beta_u^2$$

Using this

$$E = \gamma_u mc^2 \approx mc^2 + \frac{1}{2}mc^2\beta_u^2 \approx mc^2 + \frac{1}{2}mu^2$$

Hence for low speeds, the kinetic energy is $K \approx mu^2/2$. This is of course just the non-relativistic kinetic energy, which is now seen to be an approximation to a more complicated expression. The dependence of E and pc on β_u is shown in the left of Fig. 1.

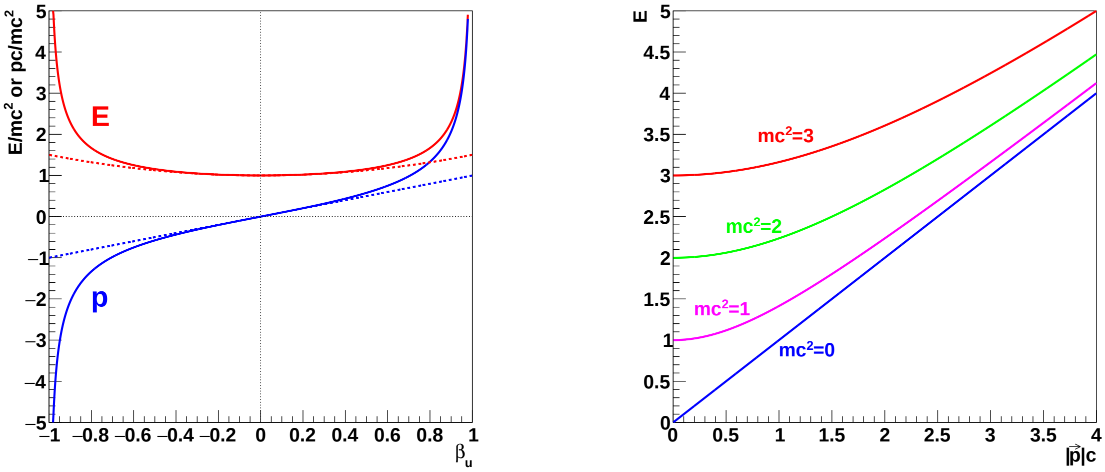

Figure 1: Left: Energy and momentum as a function of β_u. The dashed lines show the classical approximations. The energy and momentum are plotted in units of mc^2. Right: Energy as a function of momentum magnitude for various values of mass.

5 Energy-momentum invariant

Since all four-vectors transform in the same way, the length-squared for every four-vector must be an invariant. For any four-vector $(a, \vec{b})$, the length-squared is $a^2 - |\vec{b}|^2$. What is this for the energy-momentum four-vector? We can find this easily

$$E^2 - (pc)^2 = \gamma_u^2 m^2 c^4 - \gamma_u^2 \beta_u^2 m^2 c^4 = \gamma_u^2 (1 - \beta_u^2) m^2 c^4 = m^2 c^4 = (mc^2)^2$$

which is the rest energy squared. This is in fact obvious from considering the invariant in the rest frame, where $\vec{p} = 0$ and $E = E_0 = mc^2$. As mass is a real value, then $(mc^2)^2$ is always positive and you may hear the four-momentum referred to as a 'time-like' four-vector, although it has no implications for causality as it does not refer to time-ordering of events.

The above equation is often rearranged to give

$$E^2 = (pc)^2 + (mc^2)^2 = p^2c^2 + m^2c^4$$

In this form, it is clear that E is always greater than both pc and mc^2. A plot of how energy varies with momentum is shown in Fig. 1. It is clear that $E \to mc^2$ as the momentum goes to zero (i.e. the energy is just the rest energy) and it asymptotes to $E \to |\vec{p}|c$ for high momentum (i.e. the mass becomes negligible compared with the momentum). This is expanded to show the whole four-vector Lorentz transformation space in Fig. 2, which illustrates that the four-momentum is physical only in the upper central region.

6 Velocity transformation

We can now easily rederive the result on the velocity transformation which we saw in Lecture 4 using the four-momentum. For motion along the x axis, $u = pc^2/E$ and under a Lorentz transformation also along the x axis

$$E' = \gamma(E - \beta pc), \qquad p'c = \gamma(pc - \beta E)$$

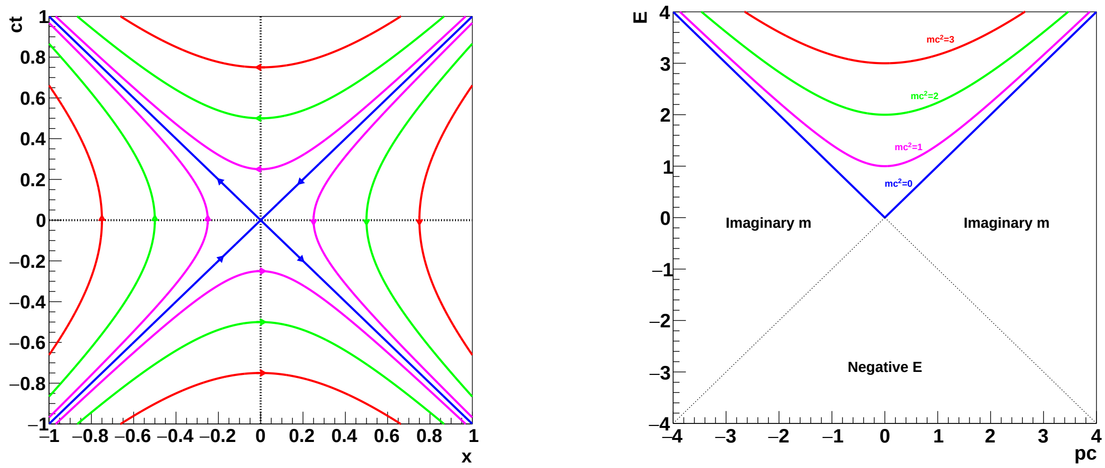

Figure 2: Left: The Lorentz transformations for ct and x. Right: Energy as a function of momentum for various values of mass expanded to show the whole four-vector Lorentz transformation space, showing why the four-momentum is unphysical in the other regions.

Hence

$$u' = \frac{p'c^2}{E'} = \frac{\gamma(pc^2 - \beta cE)}{\gamma(E - \beta pc)} = \frac{pc^2 - vE}{E - vp} = \frac{pc^2/E - v}{1 - vp/E} = \frac{u - v}{1 - vu/c^2}$$

which is, of course, the same as we got previously.

7 Photons

In the case of photons, the E and p equations need to be used carefully. We have seen no object can actually be accelerated to the speed of light because it would take infinite energy. However, photons (being light itself) do go at the speed of light. This is only possible because they are a special case in two ways; they have zero mass and always have speed c.

Putting $m = 0$ into the equations for E and p in terms of β_u seems to imply E and p would be zero. However, $|\vec{u}| = c$ and hence $|\vec{\beta}_u| = 1$, so γ_u becomes infinite. Putting this into the equations implies E and p would be infinite. The trick is to realise that m and γ_u only appear multiplied together as $\gamma_u m$. Although infinity times zero is not well-defined, you can roughly think of the product as being finite. The correct way to handle this is to eliminate this product from the equations. In fact, we already did this in writing $\vec{\beta}_u = \vec{p}c/E$. Since $|\vec{\beta}_u| = 1$ for photons, this means

$$E = |\vec{p}|c$$

which does indeed hold for a photon and is also shown in Fig. 1. In fact, this can also be found from the one equation which has m *without* a factor of γ_u, namely

$$E^2 = (pc)^2 + (mc^2)^2 = (pc)^2 \quad \text{for } m = 0$$

Note that $|\vec{u}| = c$ and $m = 0$ must *both* be true for photons because if only one were true, E and $\vec{p}$ would be zero (if $m = 0$) or infinite (if $|\vec{u}| = c$).

The above equation raises another issue. As shown in the last lecture (and is also true classically), for a particle with a given mass and velocity, the energy and momentum are fixed. We have said for all photons that $m = 0$ and speed $= c$ so do all photons have the same E and p?

No; they also vary in frequency f (or equivalently wavelength λ). This actually seen most clearly if we consider Quantum Mechanics. The Planck-Einstein relation states $E = hf$ while the de Broglie relation states $|\vec{p}| = h/\lambda$. Hence, the photon energy and momentum are determined by the light frequency (or equivalently wavelength). However, we might be worried because Quantum Mechanics is not a relativistic theory. For the case of the photon, since $E = |\vec{p}|c$, then in Quantum Mechanics this gives $hf = hc/\lambda$, which means $f\lambda = c$ as required. Hence, Relativity is consistent with Quantum Mechanics with regard to these relations.

8 Revision questions

- What is the form of the energy-momentum four-vector?
- Is it possible to accelerate an object to speed c?
- How do the properties of photons differ from the classical limit?
- What is relation between E and p for photons?

First Year Special Relativity – Lecture 8
Rest mass energy and particle decays

Mitesh Patel, 30th May 2023

1 In this lecture

- A word on units;
- The meaning of mass in Relativity;
- Look at energy and momentum conservation and the implications for mass conservation;
- Look at the kinematics of a particle decays.

2 Introduction

The material in this lecture is covered in Young and Freedman, Sec. 37.8 and in McCall in Secs. 6.3 to 6.5.

In the previous lecture, we saw that energy and momentum form a four-vector, like space and time, and hence they also undergo Lorentz transformations. In classical mechanics, energy and momentum are conserved and this is essential for calculating what happens when objects collide. The same is true relativistically. However, mass needs to be handled quite differently.

3 Units

In practice, it is very hard to get macroscopic objects to move at any significant fraction of c. Hence, the experimental tests of four-momentum in Special Relativity effectively all involve elementary particles. The particle energies involved are usually far smaller than those we experience in our everyday life. We will use units which are better suited to these scales than SI units. Energies will be measured in electron-volts (eV) which is the energy acquired by an electron (or any particle with a charge e) in accelerating through $1\,\mathrm{V}$. There are also keV ($10^3\,\mathrm{eV}$), MeV ($10^6\,\mathrm{eV}$), GeV ($10^9\,\mathrm{eV}$), etc.

The combination pc also has dimensions of energy and so can be given in the same units. Hence, p itself is energy over c, so momentum can be written in units of eV/c (or keV/c, etc.). Similarly, mc^2 can be given in eV and hence m can be given in eV/c^2. For example, in SI units the electron mass is $m_e = 9.11 \times 10^{-31}\,\mathrm{kg}$ so $m_e c^2 = 8.2 \times 10^{-14}\,\mathrm{J}$. Since $1\,\mathrm{eV} = 1.60 \times 10^{-19}\,\mathrm{J}$, then $m_e c^2 = 5.11 \times 10^5\,\mathrm{eV}$ or $0.511\,\mathrm{MeV}$. Therefore $m_e = 0.511\,\mathrm{MeV}/c^2$.

4 Implications of the rest mass energy

The existence of the rest mass energy $E_0 = mc^2$ has profound physical consequences. This equation can be considered in both directions, i.e. that energy implies mass ($m = E_0/c^2$), but also that mass implies energy ($E_0 = mc^2$). Let's consider both these in turn.

Consider a box containing some material of any type, where the total momentum of the box and material is zero, so the total system is at rest, even if not all the parts are stationary. Classically, its mass would be simply the sum of the masses $\sum_i m_i$ of the box and its contents, and so would be a constant. However, $m = E_0/c^2$ means that its mass is the total energy of everything divided by c^2, where the energy includes all the rest mass energies $m_i c^2$, but can also have contributions from any other types of energy of the constituents; kinetic, potential, thermal, binding, etc. E.g. if we heat up the box, then the material inside has more thermal energy (or equivalently, kinetic energy of the molecules it is made of) and so the total energy increases. Hence, the heated box has a higher mass than the cold box. Similarly, your phone weighs more when it is fully charged. Another example: two objects with the same sign electrostatic charge will have a higher potential energy if they are placed close together. Their total energy will be $E = m_1c^2 + m_2c^2 + V$ and so the total mass will be higher than the sum of the two masses by V/c^2. Conversely, if they have opposite sign charges, the total mass will be lower. A physical example of this is the hydrogen atom, which is made by bringing an electron and a proton together. If these two particles are initially at rest, they will bind together with a binding energy (in the lowest energy configuration) of around 13 eV. This energy is emitted as a photon when the atom is formed. Classically we would say the hydrogen atom has a mass of $m_e + m_p$. In Relativity, as both the electron and proton are initially at rest, the original system has a total energy of $(m_e + m_p)c^2$. The hydrogen atom is formed by emitting 13 eV, so the remaining energy of the hydrogen atom is $(m_e + m_p)c^2 - 13\,\mathrm{eV}$ in its rest frame. This means its mass must be $m_e + m_p - 13\,\mathrm{eV}/c^2$; it is actually less than the classical expectation by 13 eV. However, $m_p = 938.3\,\mathrm{MeV}$ and $m_e = 0.511\,\mathrm{MeV}$ so this is a correction of 13 eV in 938.8 MeV, or approximately a change of 10^{-8} of the mass. This is typical of mass changes in classical physics, which is why the assumption of the sum of masses being constant is a very good approximation for low speeds.

What about considering how mass implies energy? Given that we now know that mass can be changed, this means an increase in mass requires energy input, or conversely a decrease in mass will emit energy. In particular, if we have a reaction where the final objects have less mass than the original objects, e.g. the binding energy is increased, then the difference of the rest mass energy will be liberated. The hydrogen atom is a case in point, where the energy is emitted as a photon. However, the protons and neutrons in nuclei are held together by the strong force which has much higher binding energies. Hence, nuclear reactions can result in much larger mass changes and hence much larger energy releases. This is the basic physics underlying atomic weapons and power stations.

One final point on this topic: we talk about the electron having a mass with a very specific value, as given above. Why isn't this able to vary as well? The electron is one of what we consider to be the fundamental particles which have no internal structure. There is no way to put energy into an electron at rest; you cannot "heat" an electron. Hence, fundamental particles do have well-defined fixed values for mass.

5 Mass conservation

Energy and momentum are conserved in all processes for an isolated system; this is due to fundamental properties of time and space and holds both classically and relativistically. However, as discussed above, the mass of an object can change so the sum of the masses will not in general be conserved in reactions. This only appears to be true in classical physics because we can only rearrange the electrons and nuclei and don't change the nuclei themselves, or the numbers or types of the fundamental particles. The changes in mass are then very small and so negligible; again classical physics is an approximation to the exact relativistic theory.

However, in what initially sounds like a contradiction to the above, there *is* a mass quantity which is conserved. This is the total mass m_T defined through the equation

$$E_T^2 = p_T^2c^2 + m_T^2c^4 \qquad \text{so} \qquad m_T = \frac{\sqrt{E_T^2 - p_T^2c^2}}{c^2}$$

where E_T and $\vec{p}_T$ are the total energy and momentum. Since E_T and $\vec{p}_T$ are conserved, then clearly m_T will be as well. Note that m_T is defined using the total energy and momentum and is not the sum of the individual particle masses; $m_T \neq \sum_i m_i$. It is often the case that m_T does not correspond to the mass of any physical particle in the system.

6 Particle decay

We will consider a simple example of the above which is also an extreme case in terms of changing mass. The Higgs boson was discovered in 2012 at CERN and one way it was observed was when it decayed to two photons; $H \to \gamma\gamma$. (Note, γ is the standard symbol for a very high energy photon; do not confuse this with the Lorentz transformation parameter!) Hence, before the decay occurs, there is a Higgs boson which has a mass of $m_H = 125.2\,\mathrm{GeV}/c^2$ (which is around 130 times the proton mass). After the decay, we have two photons which both have zero mass. It is clear the initial particle mass is certainly not equal to the sum of the two final particle masses. The Higgs mass has been converted (using $E = mc^2$) into energy and this energy is carried by the photons. Consider this decay in the Higgs rest frame as shown in Fig. 1.

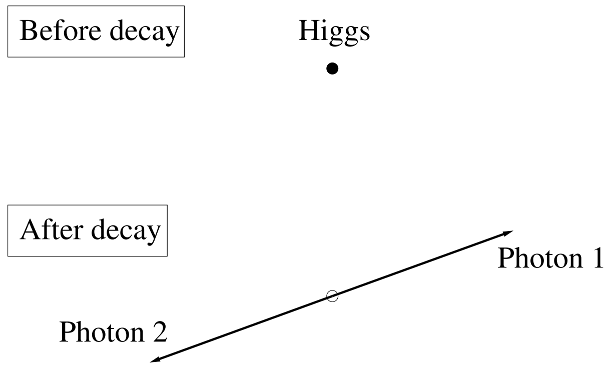

Figure 1: Particles both before and after the decay of a Higgs to two photons $H \to \gamma\gamma$ in the rest frame of the Higgs.

The total four-momentum is then just the Higgs four-momentum and so $(E_T, \vec{p}_T) = (E_H, \vec{p}_H) = (m_H c^2, \vec{0})$. To keep the total momentum zero, the two photons must be moving back-to-back with the same magnitude of momentum and hence, since $E = |\vec{p}|c$, they have the same energy E_γ. Therefore $E_T = m_H c^2 = 2E_\gamma$ so $E_\gamma = m_H c^2/2 = 62.6\,\mathrm{GeV}$. The photons equally share the energy created from the Higgs mass in this frame. The total energy ($E_T = m_H c^2$) and momentum ($\vec{p}_T = \vec{0}$) are unchanged so

$$m_T = \frac{\sqrt{E_T^2 - p_T^2 c^2}}{c^2} = m_H$$

is true both before and after the decay.

In a frame where the Higgs is moving, then the calculation can get very complicated as the photons can come off at arbitrary angles with respect to the direction of motion of the Higgs. The photon energies will not then be equal. However, the total four-momentum is still conserved, so

$$E_T = E_H = E_{\gamma 1} + E_{\gamma 2} \qquad \text{and} \qquad \vec{p}_T = \vec{p}_H = \vec{p}_{\gamma 1} + \vec{p}_{\gamma 2}$$

still holds. Clearly the invariant of the Higgs four-momentum, and hence also the total four-momentum, is $(m_T c^2)^2 = (m_H c^2)^2$ as before as invariants are the same in all frames. Critically, because the sum of two four-vectors is also a four-vector, then the invariant of the sum of the two photon four-momenta is also $(m_H c^2)^2$. This is in fact how the discovery of the Higgs was achieved. The Higgs lives for too short a time to be observed in a particle detector directly, but when it decays to two photons, these are easily detected and their energies and positions measured. By calculating the four-momentum of each photon and adding these to give the total four-momentum, then the total invariant, and hence m_T, can be found from this. This should be consistent with m_H if the two photons came from a Higgs, and indeed an excess at a mass around $125\,\mathrm{GeV}/c^2$ was found.

In the more general case, a particle can decay to two 'daughter' particles with unequal masses. For example this could be a large nucleus undergoing alpha decay to result in a smaller nucleus and the alpha particle. If the decaying particle has mass m_0 and the daughter particles have m_1 and m_2, then we need $m_0 \geq m_1 + m_2$ as otherwise there would not be enough energy to create the two daughter particles.

Again, consider the situation in the decaying particle rest frame. The initial energy is $E = m_0c^2$ and initial momentum $\vec{p} = 0$. After the decay, energy conservation requires

$$E_1 + E_2 = m_0c^2 \tag{1}$$

The two daughter particles must again be back-to-back to conserve momentum, i.e. $\vec{p}_2 = -\vec{p}_1$ so $p_2^2 = p_1^2$. Their momenta magnitudes are related to their energies by

$$p_1^2c^2 = E_1^2 - m_1^2c^4, \qquad p_2^2c^2 = E_2^2 - m_2^2c^4$$

so momentum conservation requires

$$E_1^2 - m_1^2c^4 = E_2^2 - m_2^2c^4$$

This looks hard to solve but there is a trick. Rearranging gives

$$E_1^2 - E_2^2 = (E_1 + E_2)(E_1 - E_2) = m_1^2c^4 - m_2^2c^4$$

Using Eqn. 1 this becomes

$$E_1 - E_2 = \frac{m_1^2c^4 - m_2^2c^4}{m_0c^2} \tag{2}$$

Taking the sum and difference of Eqns. 1 and 2 gives

$$E_1 = \frac{m_0c^2}{2} + \frac{m_1^2c^4 - m_2^2c^4}{2m_0c^2} = \frac{m_0^2c^4 + m_1^2c^4 - m_2^2c^4}{2m_0c^2}$$

and

$$E_2 = \frac{m_0c^2}{2} - \frac{m_1^2c^4 - m_2^2c^4}{2m_0c^2} = \frac{m_0^2c^4 + m_2^2c^4 - m_1^2c^4}{2m_0c^2}$$

Note, the equation is symmetric under interchange of $1 \leftrightarrow 2$, as would be expected. From these, the momenta can be calculated using the initial equations above but it is quite messy. Note that the particle with the higher mass gets more total energy than $m_0c^2/2$ while the lower mass particle gets less total energy. This is in direct constrast to the Galilean transformations, where the same *e.g.* momentum vector would be added each particle to boost from one frame to another. If the two decaying particles are the same type, so $m_1 = m_2$, then $E_1 = E_2 = m_0c^2/2$, as would be expected by symmetry.

The kinetic energy of the outgoing particles can in principle be accessed e.g. by letting them collide with some material and causing heating. The maximum amount of usable 'liberated' energy in the rest frame is clearly $K_1 + K_2 = (m_0 - m_1 - m_2)c^2$.

7 Revision questions

- What is the difference between the conservation of mass classically and in Relativity?
- Why can we talk about fundamental particles having a specific mass?
- What quantity is conserved in Relativistic particle decays?
- For a two-body decay such as $H \to \gamma\gamma$, can you derive the energy of the final state particles?

First Year Special Relativity – Lecture 9
Particle reactions

Mitesh Patel, 1st June 2023

1 In this lecture

- The centre-of-mass frame;
- Particle reactions.

2 Introduction

The material in this lecture is covered in Young and Freedman, Secs. 37.7 and 37.8 and in McCall in Secs. 6.3, 6.4 and 6.6.

We saw that conservation of four-momentum allowed us to work out the energy and momentum of the particles produced in the decay of a heavier particle. Here, we will look at particle reactions rather than decays; these are characterised by having two initial particles rather than just one.

3 Constant, invariant and conserved quantities

It is important to distinguish between the three concepts of a quantity being constant, invariant or conserved. These are different things.

1. A *constant* quantity has a fixed value. Examples are the speed of light c, Planck's constant h and the electron mass m_e.

2. An *invariant* is a quantity that is unchanged under a transformation. Hence a Lorentz (or four) invariant is one that is unchanged under a Lorentz transformation. However, within an inertial frame, it can change with time. Examples are any four-momentum length-squared, so $c^2\tau^2$ for the $(ct, \vec{r})$ four-position and m^2c^4 for the $(E, \vec{p}c)$ four-momentum. In fact, the equivalent of the three-vector dot-product for two four-vectors is also a Lorentz invariant (in the same way as the three-vector dot-product is a scalar under rotations). For two four-vectors $(A, \vec{a})$ and $(B, \vec{b})$, the equivalent of the dot-product is $AB - \vec{a}.\vec{b}$. Hence, the length-squared is actually the dot-product of the four-vector with itself. One example is the dot-product of the four-position and four-momentum, which is $Et - \vec{p}.\vec{r}$. You may recognise this as (the negative of) the wave phase times $\hbar$ in quantum mechanics; the phase of a wave (such as the amplitude peak) must be the same in all frames since all observers will agree on the peak.

3. A *conserved* quantity does not change with time for a given physical system. It can have a different value in another frame but its value in that frame should also be conserved. Examples are the total energy E_T and the total momentum $\vec{p}_T$ of an isolated system; these change under Lorentz transformations but whatever their value in each frame, it does not change with time.

Some quantities satisfy more than one of these labels. The speed of light c is not only constant but is also invariant (from the second postulate) and conserved. While E_T and $\vec{p}_T$ are conserved, the total mass m_T is both conserved and invariant, which is why it is a very useful quantity to work with.

4 Centre-of-mass

In the previous lecture, we calculated the particle decays in the frame where the decaying particle was at rest, as this was a convenient choice. We could then boost the result to any other frame we wanted, e.g. if the decaying particle happened to be moving in our frame.

This idea that calculations can be easier in the rest frame can be generalised to more complicated systems. Consider two particles bouncing off each other, which is the relativistic equivalent of a classical billiard ball problem. Because we now know that the total mass m_T is an important quantity, we can define the 'centre-of-mass' (CM) frame to be the one where the total momentum is zero. (This is alternatively called the 'centre-of-momentum' frame.) The total energy E_T in this frame, often labelled as $E_{\rm CM}$, is given by $E_{\rm CM} = E_T = m_T c^2$, even if there is no particle of mass m_T. Hence, $E_{\rm CM}$, despite being called an energy, is actually invariant and conserved as it is directly related to m_T.

The CM energy is important as it determines what reactions can happen, as explained later in the lecture. With $E_T^2 = p_T^2 c^2 + m_T^2 c^4$ and m_T the same in all frames, then since the momentum is non-zero in any other frame than the CM frame, E_T is higher in those frames. This higher incoming energy does not change what reactions can happen; it is used to provide the kinetic energy for the whole system which has non-zero momentum in this frame. Hence, some energy in every other frame than the CM frame is 'wasted' and not available for the reaction.

5 Particle reactions

We will consider particle reactions where two incoming particles collide and can produce some number (at least one) of outgoing particles.

Consider a photon being deflected by colliding with an electron in the CM frame; we could write the reaction as $\gamma + e \to \gamma + e$. Hence both the initial momenta $\vec{p}_{\gamma\,i} = -\vec{p}_{e\,i}$ and final momenta $\vec{p}_{\gamma\,f} = -\vec{p}_{e\,f}$ balance to give overall momentum conservation. Energy conservation requires $E_T = E_{\gamma\,i} + E_{e\,i} = E_{\gamma\,f} + E_{e\,f}$. It does not take long to convince yourself that both particles have the same initial and final momentum magnitude, and hence keep the same initial and final energy; $E_{\gamma\,i} = E_{\gamma\,f}$ and $E_{e\,i} = E_{e\,f}$. Hence all that happens is they bounce off at some different angle. Note, this is identical to classical billiard ball elastic collisions; it is basically purely set by the energy and momentum conservation. Again, the final four-momentum can be Lorentz transformed to see how this appears in other frames. This reaction is known as Compton scattering, after the person who first did the relativistic calculation and subsequent experiment to demonstrate it was correct.

We may want to calculate the actual energy and momentum needed to give a particular $E_{\rm CM}$. Most of the work for this has already been done. The trick is to 'pretend' there is a particle of just the right mass, i.e. m_T, which is created when the incoming photon and electron collide and this pretend particle then decays to give the outgoing photon and electron. Considering the decay step, then this is identical to the situation in the last lecture, where we derived the energy of the two outgoing particles in terms of the decaying particle mass. This was done in the rest frame of the decaying particle, which is the same frame as what we now call the CM frame. Since the pretend decaying particle mass is now $m_0 = E_{\rm CM}/c^2$, then this gives

$$E_1 = \frac{E_{\rm CM}^2 + m_1^2 c^4 - m_2^2 c^4}{2E_{\rm CM}} \qquad \text{and} \qquad E_2 = \frac{E_{\rm CM}^2 + m_2^2 c^4 - m_1^2 c^4}{2E_{\rm CM}}$$

where, for the example of Compton scattering, 1 could refer to the photon and 2 the electron. What about the energy of the incoming particles? This is simple once you realise that the energy and momentum conservation argument does not rely on the time order. Hence, the process of the incoming particles forming the pretend particle is identical in terms of energy and momentum conservation to the pretend particle decaying to these particles. Hence the same equations as above also hold.

However, there are more complicated cases where the particles react and different particles come out. This can change the sum of the particle masses (as happens in decays) and this change can be in either direction, so the final state could have more kinetic energy from 'destroying' mass, or the initial state kinetic energy could be absorbed into 'creating' mass. You probably know matter and antimatter annihilate each other if they come into contact. An explicit example is an electron e^- and its antiparticle, called the 'positron' e^+. These have identical masses but opposite charges, as indicated by their symbols. When these are brought together, they react and often create a pair of photons; the reaction is written as

$$e^+ + e^- \to \gamma + \gamma$$

In the CM frame, since $m_{e^+} = m_{e^-}$, then $E_{e^+} = E_{e^-} = E_{\mathrm{CM}}/2$, as can also be found from the general equations with $m_1 = m_2$. Since the photons have no mass, then to have equal and opposite momenta they must each have the same energy, so $E_{e^+} = E_{e^-} = E_\gamma$ and these are all equal to $E_{\mathrm{CM}}/2$. This is therefore a very simple case, due to having equal mass particles in both cases.

If we think in terms of the masses, then the initial state has a sum of particle masses of $2m_e$, while the final state particle mass sum is clearly zero. Therefore, this is a case where the initial masses are destroyed and the rest mass energy has been released, here going into the energy of the photons. This means the reaction can proceed even if the electron and positron have negligible kinetic energy, i.e. are effectively at rest before they touch each other. In the case they are moving very slowly then $E_{\mathrm{CM}} \approx 2m_ec^2$ and hence $E_\gamma \approx m_ec^2 \approx 0.511\,\mathrm{MeV}$. This reaction would therefore result in a particular line in the photon spectrum at the value of the electron rest mass energy. Such a line has been detected by gamma ray astronomy observations from sources such as black holes, where positrons can be created through other reactions before then annihilating with electrons and giving photons. However, this spectral line has also been used to search for antimatter in other parts of the galaxy. If there are antimatter stars, then there must be some boundary in the galaxy between the matter and antimatter parts. Interstellar gas (which does not have relativistic speeds) along that boundary would be a mix of matter and antimatter and so would be undergoing the above reaction. No such signals of annihilation along boundaries have been seen, so we believe the galaxy is purely made of matter.

6 Reaction thresholds

Another reaction which can occur is when the electron and positron annihilation produces a muon and its antiparticle, called an antimuon.

$$e^+ + e^- \to \mu^+ + \mu^-$$

The muon is effectively identical in its properties to the electron but has a much larger mass; $m_\mu c^2 = 106\,\mathrm{MeV}$ so $m_\mu/m_e = 207$. Hence, this is a case where the initial sum of particle masses $2m_e$ is less than the final sum of particle masses $2m_\mu$. This means kinetic energy in the initial particles is needed to form the final particles. By an identical argument to the photon case, we have $E_{e^+} = E_{e^-} = E_{\mu^+} = E_{\mu^-}$. However, the minimum value of E_{μ^+} and E_{μ^-} must each be $m_\mu c^2$, even if the muons have no kinetic energy. This means unless $E_{\mathrm{CM}} \geq 2m_\mu c^2$, the reaction cannot occur. There is a 'threshold' for the electron energies required for this reaction to occur, in contrast to the previous photon reaction where the reaction can always occur for any E_{CM}. Specifically, if $E_{e^+} = E_{e^-} < m_\mu c^2$, then there is not enough initial energy to make the mass of the pair of muons at all.

If we set the electron energies to be just at the threshold, so $E_{\rm CM} = 2m_\mu c^2$, then there is just enough energy to create the muons but they will not have any kinetic energy. Hence, the muons will be created at rest. If electrons with more energy than this minimum are used, the energy above threshold becomes muon kinetic energy.

It should be clear that observers in any inertial frame will agree on whether muons are produced or not. As discussed previously, all other inertial frames than the CM frame will have a higher total energy. However, no matter what the energy in any frame, if $E_{\rm CM} < 2m_\mu c^2$ then despite the higher energy, the reaction can still not occur. This is why $E_{\rm CM}$ (or equivalently m_T) is the critical energy value for a reaction, not the total E in any arbitrary observer frame.

The idea of a threshold is very general and does not just apply to cases with two final particles. If a reaction creates N final particles, then if the sum of all the final particle masses is greater than the sum of the two initial particle masses, there will be a energy threshold for the reaction to occur. The threshold will be $E_{\rm CM} = \sum_i^N m_i c^2$ and if exactly at this threshold, all the final particles will be at rest, while above this threshold, they will be moving and so have kinetic energy. There is one important special case, which is $N = 1$, i.e. creating a single particle. As before, if the mass M of this particle is more than the sum of the masses of the incoming particles, there will be a threshold of $E_{\rm CM} = Mc^2$. However, the single particle must have no momentum in the CM frame, as momentum is conserved. Therefore, if the incoming energy is raised above the threshold, the particle created cannot simply move to gain kinetic energy. This means a single particle can only be created if $E_{\rm CM} = Mc^2$ with the reaction not happening for lower *or* higher energies. Hence, unless the incoming particle energies are set precisely to the right values, production of the single particle cannot occur. This type of reaction is called a 'resonance' because, in quantum mechanics, it is effectively the same mathematics as for a resonance in a harmonic oscillator.

The above discussion explains why modern high energy particle accelerators like the Large Hadron Collider at CERN in Geneva are built the way they are. Firstly, they use two beams which are collided head-on, so all the energy goes into $E_{\rm CM}$ and there is no 'wasted' kinetic energy for the overall system. Secondly, the high beam energies give a high $E_{\rm CM}$ and so are more likely to be above threshold for creating previously undiscovered heavy new particles. This is precisely how the Higgs was able to be created and hence discovered.

7 Revision questions

- What is the centre-of-mass frame and how can you find such a frame?
- Why is this frame useful when trying to solve for kinematic quantities?
- What dictates whether reactions can occur in the case when the final state particles are heavier than the initial state particles?

First Year Special Relativity – Lecture 10
The relativistic Doppler effect

Mitesh Patel, 2nd June 2023

1 In this lecture

- The relativistic Doppler effect.

2 Introduction

The material in this lecture is covered in Young and Freedman Sec. 37.6 and in McCall in Sec. 6.7.

You will know about the Doppler effect in sound; it is pretty obvious whenever an ambulance passes you in the street. Waves emitted from an object approaching you get bunched up and so have a shorter wavelength, giving a higher frequency, while the opposite is true as the object moves away from you. The same effect also applies to light. However, for objects moving at relativistic speeds, there is an additional effect due to time dilation which needs to be taken into account first.

3 Time dilation

Consider a relativistic object emitting light where the light is travelling purely perpendicular to the direction of motion of the object. Classically there would be no Doppler shift. This is equivalent to the moment the ambulance is closest to you; at that time you hear the "true" frequency.

However, we know that time for a moving object is dilated, and so is slowed relative to the observer. Hence, if the object emitted light of a frequency f in its rest frame, it will actually emit light of frequency $f_d = f/\gamma_u$ when moving perpendicular to an observer and hence the observer will always measure a lower frequency in this orientation. It terms of colours, this light is moved from the blue end of the spectrum towards the red end, and this is called 'red-shifted' light, as opposed to 'blue-shifted' which is when light is pushed higher in frequency.

4 The stretch or squeeze of wavelengths

We now need to consider the effect of the wavelength of light being squeezed or stretched. This is effectively just geometry and so also holds relativistically. The frequency of light observed depends on whether the observer is in front (or forward) of the source in its direction of motion, or behind (or backward), as shown in Fig. 1

Specifically, assume the light source is moving with speed u along the x axis. Consider one full wavelength of light also emitted in the $+x$ direction as shown in Fig. 2. Emitting one wavelength will take a time equal to one period $T = 1/f_d$; note we must use f_d as this is the source frequency in the observer frame.

During this time, the leading part of this wave will have gone a distance cT while the emitting object will have gone uT. Hence the wave for one period will be squashed into a length $(c-u)T$, i.e. this is its wavelength. Hence

$$\lambda_f = (c-u)T = \frac{c-u}{f_d} = \frac{c\gamma_u(1-\beta_u)}{f} = \gamma_u(1-\beta_u)\lambda$$

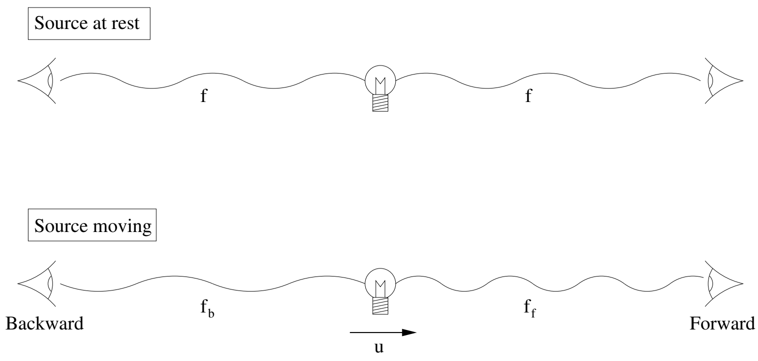

Figure 1: A moving source gives different frequencies to a forward (f) and backward (b) observer.

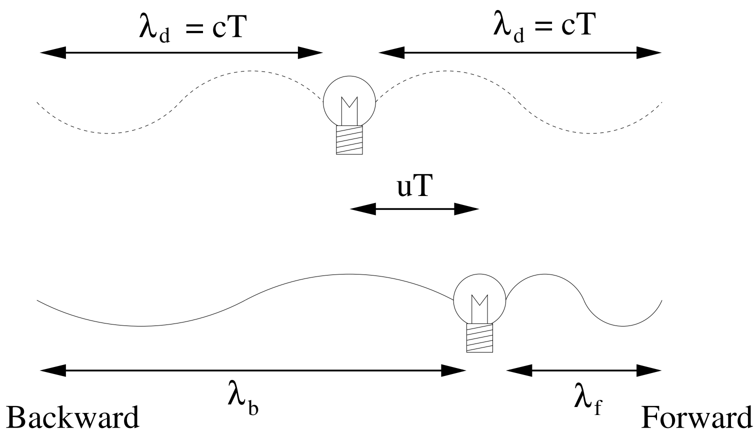

Figure 2: Emission of one wavelength of light for a source moving with speed u. Top: The dashed lines show the wave which would have been emitted if the source had not been moving to the right. Bottom: The actual waves emitted by the moving source.

where the subscript f indicates this is as seen by an observer in front of the moving source. Its frequency will therefore be

$$f_f = \frac{c}{\lambda_f} = \frac{c}{\gamma_u(1-\beta_u)\lambda} = \frac{f}{\gamma_u(1-\beta_u)}$$

If the light is going in the $-x$ direction, then the wave from one period will fit into a length $(c+u)T$ so

$$\lambda_b = \gamma_u(1+\beta_u)\lambda \qquad \text{and} \qquad f_b = \frac{f}{\gamma_u(1+\beta_u)}$$

where b indicates backwards, i.e. the observer is behind the moving source. Note, these are exactly the same equations with β_u replaced by $-\beta_u$, as might be expected.

There is another common formula for the above, using the explicit form for γ_u. Consider

$$\gamma_u(1-\beta_u) = \frac{1-\beta_u}{\sqrt{1-\beta_u^2}} = \sqrt{\frac{(1-\beta_u)^2}{(1-\beta_u)(1+\beta_u)}} = \sqrt{\frac{1-\beta_u}{1+\beta_u}}$$

Similarly

$$\gamma_u(1+\beta_u) = \frac{1+\beta_u}{\sqrt{1-\beta_u^2}} = \sqrt{\frac{(1+\beta_u)^2}{(1-\beta_u)(1+\beta_u)}} = \sqrt{\frac{1+\beta_u}{1-\beta_u}}$$

Note, this tells us that

$$\gamma_u(1-\beta_u) = \frac{1}{\gamma_u(1+\beta_u)}$$

which is easy to verify since it means

$$\gamma_u^2 = \frac{1}{(1-\beta_u)(1+\beta_u)} = \frac{1}{1-\beta_u^2}$$

as expected. Note that these two factors, which are the inverse of each other, are also simply related by $\beta_u \to -\beta_u$. Hence, we can write

$$\lambda_f = \lambda\sqrt{\frac{1-\beta_u}{1+\beta_u}} \qquad \text{and} \qquad f_f = f\sqrt{\frac{1+\beta_u}{1-\beta_u}}$$

while

$$\lambda_b = \lambda\sqrt{\frac{1+\beta_u}{1-\beta_u}} \qquad \text{and} \qquad f_b = f\sqrt{\frac{1-\beta_u}{1+\beta_u}}$$

In this form, it is straightforward to see that, with β_u positive as defined in setting up the problem, then $f_f > f$, i.e. this is always a blue-shift, while $f_b < f$, i.e. this is always a red-shift. The fact that $f_f > f$ means the wavelength squeeze always overcomes the time dilation effect.

Hence, it is very important to be sure of the sign of β_u and what it means, as it is easy to misinterpret a Doppler equation. It is always best to think about whether physically the light wavelength will be squeezed (blue-shifted), when the source moves towards the observer, or stretched (red-shifted), when it is moving away. This will normally ensure that you get the right sign for β_u.

We know the first postulate of Relativity states all inertial frames are equivalent. Hence, the above, which relate to a moving source and stationary observer, must also hold (from the observer's perspective) in another inertial frame where the source is stationary and the observer is moving towards or away from the source. Note that this is very different from sound. In that case, you have to specify whether the source or observer (or both) are moving because the sound moves at a fixed speed relative to the air. In this respect, the relativistic Doppler shift is simpler as light always goes at speed c and all inertial frames are equivalent so it doesn't matter whether the source or the observer are moving; in fact we cannot even define this absolutely because, as always, it is purely a matter of which inertial frame the source and observer are in.

As you will know, the Universe is expanding, which results in galaxies further from us on average appearing to move away faster than ones closer to us. The speeds the galaxies are moving away can be measured by the (usually) red shift of specific lines in spectra from atoms. While part of this effect is related to the Doppler shift discussed above, some of it is also a General Relativity effect of the actual space between us and the other galaxies being stretched by the expansion itself.

5 Doppler shift from four-momentum

We can consider the Doppler effect from a different perspective by considering it in terms of photons. Consider a stationary tank firing a shell which has speed u. Classically, if the tank starts moving with speed v in the direction it is firing, the next shell will have speed $u + v$ and so will have a higher momentum and hence higher kinetic energy. The shell is "thrown forwards" by the tank's motion and photons work in exactly the same way. Although we cannot speed up photons (as they always go at the speed of light), a moving source can still give extra momentum and hence energy to photons emitted in the same direction. Therefore, let's try to find the Doppler shift result a different way using the fact that we know about the four-momentum and quantum mechanics.

Consider a photon emitted in the source rest frame along the $+x$ axis. The light has frequency f and hence the photons in the light have energy $E = hf$ and a momentum $pc = hc/\lambda = hf$ also. If we do a Lorentz transformation, we can move to a frame where the source is moving. As shown before, when we do a Lorentz transformation with speed v on a object at rest (here the source), the object speed is $u = -v$ in the second frame. This means that to have β_u positive, as defined when setting up the problem, then we need to Lorentz transform in the negative direction. Transforming the photon then gives

$$E' = \gamma(E - \beta pc) = \gamma_u(E + \beta_u pc) = \gamma_u(hf + \beta_u hf) = hf\gamma_u(1 + \beta_u)$$

and since $E' = hf_f$

$$f_f = f\gamma_u(1 + \beta_u) = \frac{f}{\gamma_u(1 - \beta_u)}$$

which is the same result as previously. The above assumed pc was positive; putting it negative would be equivalent to the photon travelling in the other direction and so would give f_b.

Note this is often by far the easiest way to calculate the Doppler shift, particularly for photons emitted at arbitrary angles. However, this method doesn't clearly bring out the two different physical effects which contribute; namely time dilation and the squeeze/stretch of the wavelengths. Importantly, the Lorentz transformation method is also usable for particles with non-zero mass; for example, the Doppler shift for the quantum wavelength of a moving electron can be easily found using four-vectors.

There is one subtlety which might be overlooked in this calculation; it only depends on the photon transformation and actually has nothing to do with the source motion. The calculation would be identical if the source was stationary or moving in the original frame; $-\beta$ is then just the transformation parameter between frames, not the final speed of the source. This is effectively again saying that all frames are equivalent, as discussed above.

We might also want to check the case considered in Sec. 3, i.e. when the motion is transverse to the photon and the only effect is time dilation. For this case, the photon must be emitted at 90° to the source motion but this has to be in the *observer* frame, not the source rest frame. This means that, while the photon must have a non-zero $p'_y = p_y$, we need $p'_x = 0$. This requires the photon to have a specific p_x such that the Lorentz transformation reduces it to zero, and this specific value is given by

$$p'_x c = \gamma(p_x c - \beta E) = \gamma_u(p_x c + \beta_u E) = 0 \qquad \text{so} \qquad p_x c = -\beta_u E$$

Hence

$$E' = \gamma(E - \beta p_x c) = \gamma_u(E + \beta_u p_x c) = \gamma_u(E - \beta_u^2 E) = \frac{E}{\gamma_u}$$

and so, since $E' = hf_d$ in the previous notation, then $f_d = f/\gamma_u$ as stated previously.

6 Revision questions

- With a light source moving in a direction perpendicular to an observer, how is the emission frequency related to the observed frequency? What physical effects contribute to this result?
- With a light source moving in a direction parallel and towards or away from an observer, how is the emission frequency related to the observed frequency? Again, what physical effects contribute to this result?
- How are the direction of light emission in the rest frame and a moving frame related?