Special Relativity

Edition 1.0 20/02/06

Note: current version of this book can be found at

http://en.wikibooks.org/wiki/Special_relativity

Table of Contents

The principle of relativity

Introduction

Special relativity (SR) or the 'special theory of relativity' was discovered by Albert Einstein and first published in 1905 in the article "On the Electrodynamics of Moving Bodies". It replaced Newtonian notions of space and time and it incorporates Maxwell's theory of electromagnetism. The theory is called "special" because it applies the principle of relativity to the "restricted" or "special" case of inertial reference frames in 'flat' spacetime where the effects of gravity can be ignored. Ten years later, Einstein published his general theory of relativity (general relativity, "GR") which incorporated these effects.

Beginners often believe that special relativity is only about objects that are moving at high velocities. This is a mistake. Special relativity applies at all velocities but at low velocity the predictions of special relativity are almost identical to those of the Newtonian empirical formulae. Special relativity introduces a deeper understanding of why physical events happen.

This book is intended for undergraduates but can be used by anyone with a higher school level of mathematics. It is arranged in two sections, a general description and a mathematical description of the theory. As a "Wikibook" it is not complete and the next edition can be edited by anyone who feels they have spotted a mistake or wishes to add more detail and clarity.

The principle of relativity

Principles of relativity address the problem of how events that occur in one place are observed from another place. This problem has been a difficult theoretical challenge since the earliest times.

Aristotle argued in his "Physics" that things must either be moved or be at rest. According to Aristotle, on the basis of complex and interesting arguments about the possibility of a 'void', things cannot remain in a state of motion without something moving them. As a result Aristotle proposed that objects would stop entirely in empty space.

Galileo challenged this idea of movement being due to a continuous action of something that causes the movement. In his " Dialogue Concerning the Two Chief World Systems" he considers observations of motion made by people inside a ship who could not see the outside:

"have the ship proceed with any speed you like, so long as the motion is uniform and not fluctuating this way and that. You will discover not the least change in all the effects named, nor could you tell from any of them whether the ship was moving or standing still. "

According to Galileo, if the ship moves smoothly someone inside it would be unable to determine whether they are moving. This concept leads to **Galilean Relativity** in which it is held that things continue in a state of motion unless acted upon.

Galilean Relativity contains two important principles: firstly it is impossible to determine who is actually at rest and secondly things continue in uniform motion unless acted upon. The second principle is known as Galileo's Law of Inertia or Newton's First Law of Motion.

Reference:

Galileo Galilei (1632). Dialogues Concerning the Two Chief World Systems. Aristotle (350BC). Physics. http://classics.mit.edu/Aristotle/physics.html

Frames of reference, events and transformations

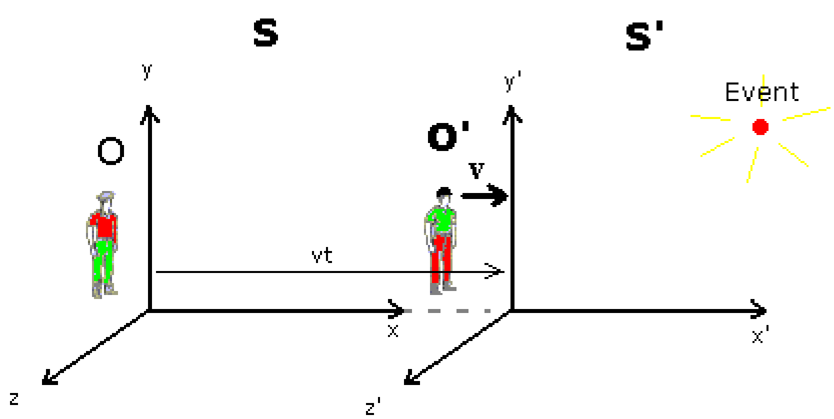
Physical observers are considered to be surrounded by a **reference frame** which is a set of coordinate axes in terms of which position or movement may be specified or with reference to which physical laws may be mathematically stated.

An **inertial reference frame** is a collection of objects that have no net motion relative to each other. It is a coordinate system defined by the non-accelerated motion of objects with a common direction and speed.

An **event** is something that happens independently of the reference frame that might be used to describe it. Turning on a light or the collision of two objects would constitute an event.

Suppose there is a small event, such as a light being turned on, that is at coordinates x,y,z,t in one reference frame. What coordinates would another observer, in another reference frame moving relative to the first at velocity v assign to the event? This problem is illustrated below:

Transformation of Coordinates



The observers are moving at a relative velocity of v and each observer has their own set of coordinates (x,y,z,t) and (x',y',z',t'). What coordinates do they assign to the event?

What we are seeking is the relationship between the second observer's coordinates x',y',z',t' and the first observer's coordinates x,y,z,t. According to Newtonian Relativity:

$x' = x - vt$

$y' = y$

$z' = z$

$t' = t$

This set of equations is known as a **Galilean coordinate transformation** or **Galilean transformation**. These equations show how the position of an event in one reference frame is related to the position of an event in another reference frame. But what happens if the event is something that is moving? How do velocities transform from one frame to another?

The calculation of velocities depends on Newton's formula: $v = dx / dt$. The use of Newtonian physics to calculate velocities and other physical variables has led to Galilean Relativity being called **Newtonian Relativity** in the case where conclusions are drawn beyond simple changes in coordinates. The velocity transformations for the velocities in the three directions in space are, according to Galilean relativity:

$$\mathbf{u}_{\mathbf{x}}' = \mathbf{u}_{\mathbf{x}} - \mathbf{v}$$

$$\mathbf{u}_{\mathbf{y}}' = \mathbf{u}_{\mathbf{y}}$$

$$\mathbf{u}_{\mathbf{z}}' = \mathbf{u}_{\mathbf{z}}$$

This result is known as the **classical velocity addition theorem** and summarises the transformation of velocities between two Galilean frames of reference. It means that the velocities of projectiles must be determined relative to the velocity of the source and destination of the projectile. For example, if a sailor throws a stone at 10 km/hr from Galileo's ship which is moving towards shore at 5 km/hr then the stone will be moving at 15 km/hr when it hits the shore.

In Newtonian Relativity the geometry of space is assumed to be Euclidian and the measurement of time is assumed to be the same for all observers.

The derivation of the classical velocity addition theorem is as follows:

If the Galilean transformations are differentiated with respect to time:

$x' = x - vt$

So:

$dx' / dt = dx / dt - v$

But in Galilean relativity $t' = t$ and so $dx' / dt' = dx' / dt$ therefore:

$dx' / dt' = dx / dt - v$

$dy' / dt' = dy / dt$

$dz' / dt' = dy / dt$

If we write $u_x^{'} = dx^{'}/dt^{'}$ etc. then:

$$u_x^{'} = u_x - v$$

$$u_y^{'} = u_y$$

$$u_z^{'} = u_z$$

Special relativity

In the nineteenth century James Clerk Maxwell discovered the equations that describe the propagation of electromagnetic waves such as light. If one assumes that both the Maxwell equations are valid, and that the Galilean transformation is the appropriate transformation, then it should be possible to measure velocity absolutely and there should be a **preferred reference frame**. The preferred reference frame could be considered the true zero point to which all velocity measurements could be referred.

Special relativity restored a principle of relativity in physics by maintaining that although Maxwell's equations are correct Galilean relativity is wrong: there is no preferred reference frame. Special relativity brought back the interpretation that in all inertial reference frames the same physics is going on and there is no phenomenon that would allow an observer to pinpoint a zero point of velocity. Einstein extended the principle of relativity by proposing that the laws of physics are the same regardless of inertial frame of reference. According to Einstein, whether you are in the hold of Galileo's ship or in the cargo bay of a space ship going at a large fraction of the speed of light the laws of physics will be the same.

The postulates of special relativity

1. First postulate: the principle of relativity

Observation of physical phenomena by more than one inertial observer must result in agreement between the observers as to the nature of reality. Or, the nature of the universe must not change for an observer if their inertial state changes. Every physical theory should look the same mathematically to every inertial observer. Formally: **the laws of physics are the same regardless of inertial frame of reference**.

2. Second postulate: invariance of the speed of light

The speed of light in vacuum, commonly denoted c, is the same to all inertial observers, is the same in all directions, and does not depend on the velocity of the object emitting the light. Formally: **the speed of light in free space is a constant in all inertial frames of reference.**

Using these postulates Einstein was able to calculate how the observation of events depends upon the relative velocity of observers. He was then able to construct a theory of physics that led to predictions such as the equivalence of mass and energy and early quantum theory.

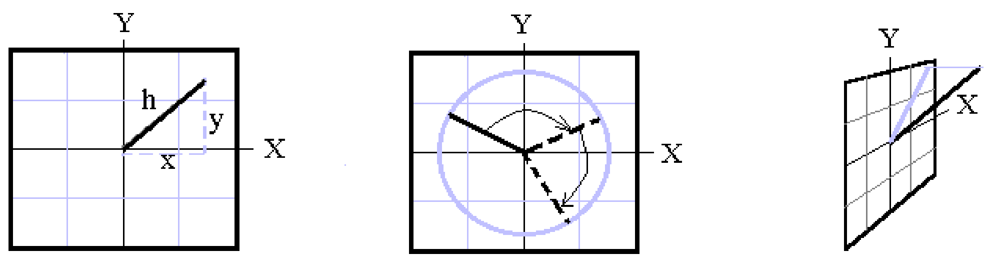
Spacetime

The spacetime interpretation of special relativity

Although the special theory of relativity was first proposed by Einstein in 1905, the modern approach to the theory depends upon the concept of a four-dimensional universe, that was first proposed by Hermann Minkowski in 1908, and further developed as a result of the contributions of Emmy Noether. This approach uses the concept of invariance to explore the types of coordinate systems that are required to provide a full physical description of the location and extent of things.

The modern theory of special relativity begins with the concept of "length". In everyday experience, it seems that the length of objects remains the same no matter how they are rotated or moved from place to place. We think that the simple length of a thing is "invariant". However, as is shown in the illustrations below, what we are actually suggesting is that length seems to be invariant in a three-dimensional coordinate system.

Figure 1: Invariance of length on a Euclidean plane.



The length of a thing in a two dimensional coordinate system is given by Pythagoras' theorem. $h^2 = x^2 + y^2$

In the 2D plane length is invariant during rotations on the plane. The length is also invariant if the thing is just moved from place to place (translational invariance).

If a thing rotates out of the plane the length it projects on the plane is no longer equal to the real length of the thing.

The length of a thing in a two-dimensional coordinate system is given by Pythagoras's theorem:

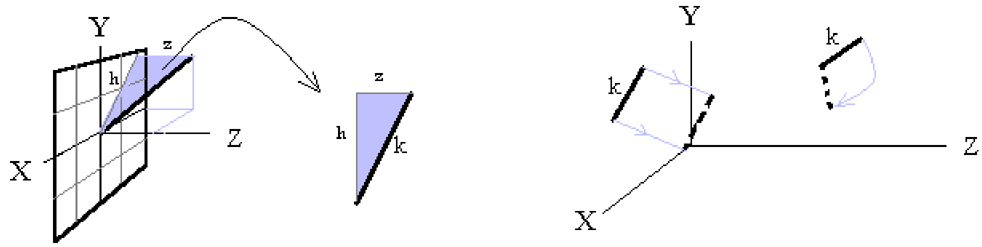
$$h^2 = x^2 + y^2$$

This two-dimensional length is not invariant if the thing is tilted out of the two-dimensional plane. In everyday life, a three-dimensional coordinate system seems to describe the length fully. The length is given by the three-dimensional version of Pythagoras's theorem:

$$h^2 = x^2 + y^2 + z^2$$

The derivation of this formula is shown in the illustration below.

Figure 2: Invariance in a 3D Euclidean space.



The length of an object in a three dimensional coordinate system is given by the 3D version of Pythagoras' theorem:

$k^2 = h^2 + z^2$ but $h^2 = x^2 + y^2$

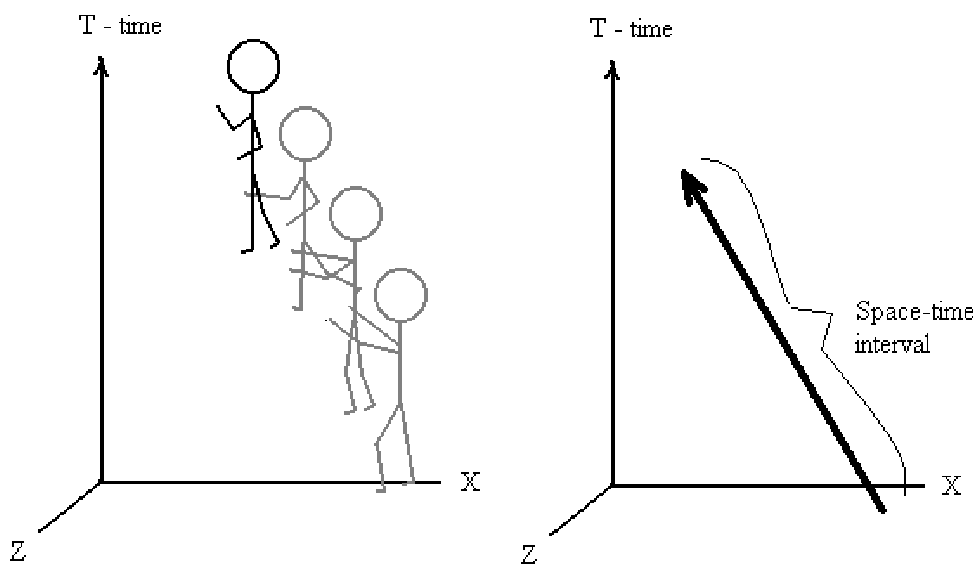
$k^2 = x^2 + y^2 + z^2$

In a three dimensional coordinate system it seems that the real length of a thing stays the same (is INVARIANT) during translations and rotations. It appears to be always given by:

$k^2 = x^2 + y^2 + z^2$

It seems that, provided all the directions in which a thing can be tilted or arranged are represented within a coordinate system, then the coordinate system can fully represent the length of a thing. However, it is clear that things may also be changed over a period of time. We must think of time as another direction in which things can be arranged. This is shown in the following diagram:

Figure 3: The invariant space-time interval.



Motions can be represented as lengths spanning both space and time in a coordinate system. These lengths are called SPACE-TIME INTERVALS. Time can be considered to be yet another direction for arranging things. This suggests that the universe could be four dimensional. If the universe is truly four dimensional then space-time intervals would be invariant when things move .

The path taken by a thing in both space and time is known as the space-time interval.

Hermann Minkowski realised in 1908 that if things could be rearranged in time, then the universe might be four-dimensional. He boldly suggested that Einstein's recently-discovered theory of Special Relativity was a consequence of this four-dimensional universe. He proposed that the space-time interval might be related to space and time by Pythagoras' theorem in four dimensions:

$s^2 = x^2 + y^2 + z^2 + (ict)^2$

Where i is the imaginary unit (sometimes imprecisely called $\sqrt{-1}$), c is a constant, and t is the time interval spanned by the space-time interval, s. The symbols x, y and z represent displacements in space along the corresponding axes. In this equation, the 'second' becomes just another unit of length. In the same way as centimetres and inches are both units of length related by centimetres = 'conversion constant' times inches, metres and seconds are related by metres = 'conversion constant' times seconds. The conversion constant, c has a value of about 300,000,000 meters per second. Now i^2 is equal to minus one, so the space-time interval is given by:

$s^2 = x^2 + y^2 + z^2 - (ct)^2$

Minkowski's use of the imaginary unit has been superseded by the use of advanced geometry, that uses a tool known as the "metric tensor", but his original equation survives, and the space-time interval is still given by:

$s^2 = x^2 + y^2 + z^2 - (ct)^2$

Space-time intervals are difficult to imagine; they extend between one place and time and another place and time, so the velocity of the thing that travels along the interval is already determined for a given observer.

If the universe is four-dimensional, then the space-time interval will be invariant, rather than spatial length. Whoever measures a particular space-time interval will get the same value, no matter how fast they are travelling. The invariance of the space-time interval has some dramatic consequences.

The first consequence is the prediction that if a thing is travelling at a velocity of c metres per second, then all observers, no matter how fast they are travelling, will measure the same velocity for the thing. The velocity c will be a universal constant. This is explained below.

When an object is travelling at c, the space time interval is **zero**, this is shown below:

The space-time interval is $s^2 = x^2 + y^2 + z^2 - (ct)^2$

The distance travelled by an object moving at velocity v in the x direction for t seconds is:

$x = vt$

If there is no motion in the y or z directions the space-time interval is $s^2 = x^2 + 0 + 0 - (ct)^2$

So: $s^2 = (vt)^2 - (ct)^2$

But when the velocity v equals c:

$s^2 = (ct)^2 - (ct)^2$

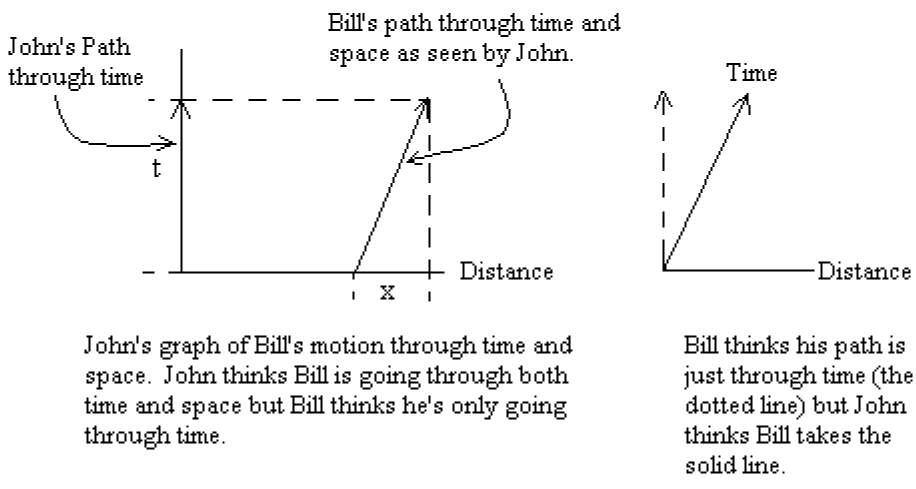
And hence the space time interval $s^2 = (ct)^2 - (ct)^2 = 0$

A space-time interval of zero only occurs when the velocity is c. When observers observe something with a space-time interval of zero, they all observe it to have a velocity of c, no matter how fast they are moving themselves.

The universal constant, c, is known for historical reasons as the "speed of light". In the first decade or two after the formulation of Minkowski's approach many physicists, although supporting Special Relativity, expected that light might not travel at exactly c, but might travel at very nearly c. There are now few physicists who believe that light does not propagate at c.

The second consequence of the invariance of the space-time interval is that clocks will appear to go slower on objects that are moving relative to you. Suppose there are two people, Bill and John, on separate planets that are moving away from each other. John draws a graph of Bill's motion through space and time. This is shown in the illustration below:

Figure 4: John and Bill - observers moving away from each other.



Being on planets, both Bill and John think they are stationary, and just moving through time. John spots that Bill is moving through what John calls space, as well as time, when Bill thinks he is moving through time alone. Bill would also draw the same conclusion about John's motion. To John, it is as if Bill's time axis is leaning over in the direction of travel and to Bill, it is as if John's time axis leans over.

John calculates the length of Bill's space-time interval as:

$s^2 = (vt)^2 - (ct)^2$

whereas Bill doesn't think he has travelled in space, so writes:

$s^2 = (0)^2 - (cT)^2$

The space-time interval, s^2, is invariant. It has the same value for all observers, no matter who measures it or how they are moving in a straight line. Bill's s^2 equals John's s^2 so:

$(0)^2 - (cT)^2 = (vt)^2 - (ct)^2$

and

$- (cT)^2 = (vt)^2 - (ct)^2$

hence

$$t = T/\sqrt{1 - v^2/c^2}.$$

So, if John sees Bill measure a time interval of 1 second ($T = 1$) between two ticks of a clock that is at rest in Bill's frame (modelled by the condition $X = 0$), John will find that his own clock measures between these same ticks an interval t, called **coordinate time**, which is greater than one second. It is said that clocks in motion slow down, relative to those on observers at rest. This is known as "relativistic time dilation of a moving clock". The time that is measured in the rest frame of the clock (in Bill's frame) is called the **proper time** of the clock.

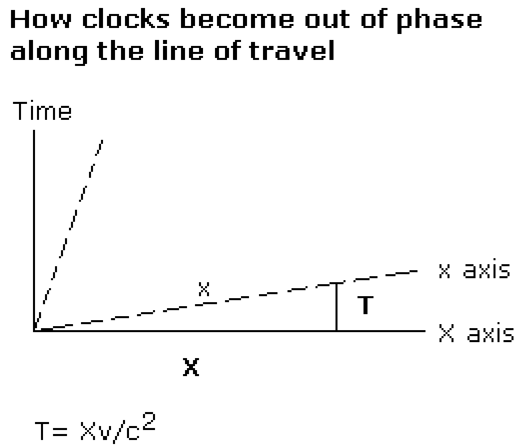
John will also observe measuring rods at rest on Bill's planet to be shorter than his own measuring rods, in the direction of motion. This is a prediction known as "relativistic length contraction of a moving rod". If the length of a rod at rest on Bill's planet is X, then we call this quantity the proper length of the rod. The length x of that same rod as measured on John's planet, is called **coordinate length**, and given by

$$x = X\sqrt{1 - v^2/c^2}.$$

See section on the Lorentz transformation below.

The last consequence is that clocks will appear to be out of phase with each other along the length of a moving object. This means that if one observer sets up a line of clocks that are all synchronised so they all read the same time, then another observer who is moving along the line at high speed will see the clocks all reading different times. In other words observers who are moving relative to each other see different events as **simultaneous**. This effect is known as **Relativistic Phase** or the **Relativity of Simultaneity**. Relativistic phase is often overlooked by students of Special Relativity, but if it is understood then phenomena such as the twin paradox are easier to understand.

The way that clocks go out of phase along the line of travel can be calculated from the concepts of the invariance of the space-time interval and length contraction.



The relationship for comparing lengths in the direction of travel is given by:

$$x = X\sqrt{1 - v^2/c^2}.$$

So distances between two points according to Bill are simple lengths in space (X) whereas John sees Bill's measurement of distance as a combination of a distance (x) and a time interval:

$x^2 = X^2 - (cT)^2$

But from : $x = X\sqrt{1 - v^2/c^2}$.

$x^2 = X^2 - (v^2 / c^2)X^2$

So: $(cT)^2 = (v^2 / c^2)X^2$

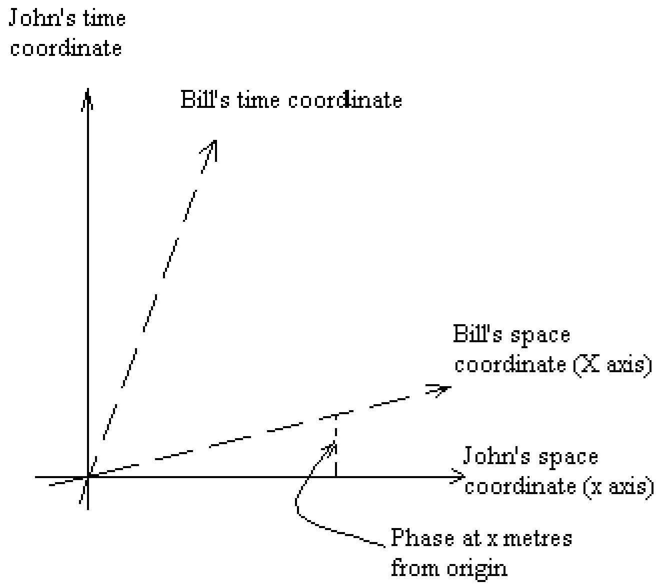
And $cT = (v / c)X$

So: $\boldsymbol{T = (v / c^2)X}$

Clocks that are synchronised for one observer go out of phase along the line of travel for another observer moving at v metres per second by :(v / c^2) seconds for every metre. This is one of the most important results of Special Relativity and is often neglected by students.

The net effect of the four-dimensional universe is that observers who are in motion relative to you seem to have time coordinates that lean over in the direction of motion, and consider things to be simultaneous, that are not simultaneous for you. Spatial lengths in the direction of travel are shortened, because they tip upwards and downwards, relative to the time axis in the direction of travel, akin to a rotation out of three-dimensional space.

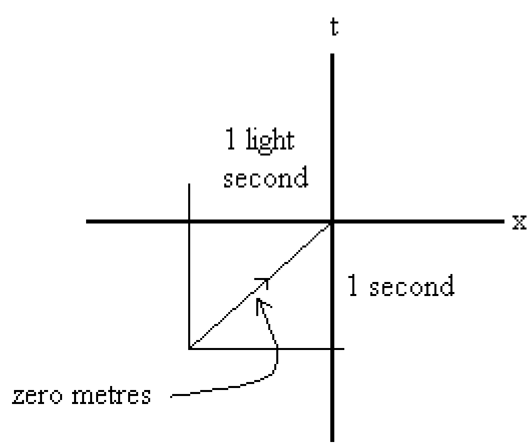
Figure 5: How Bill's coordinates appear to John at the instant Bill passes him.



How John views Bill's coordinate system

Great care is needed when interpreting space-time diagrams. Diagrams present data in two dimensions, and cannot show faithfully how, for instance, a zero length space-time interval appears.

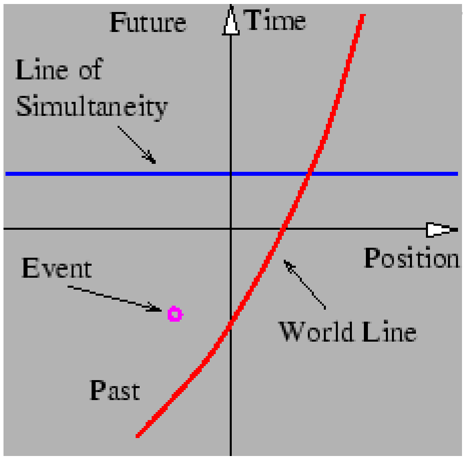
Figure 6: Space-time diagrams are often misleading



The path of a light ray is a space-time interval of zero but appears as a long line in the diagram.

Spacetime

Spacetime diagram showing an *event*, a *world line*, and a *line of simultaneity*



In order to gain an understanding of both Galilean and Special Relativity it is important to begin thinking of space and time as being different dimensions of a four-dimensional vector space called spacetime. Actually, since we can't visualize four dimensions very well, it is easiest to start with only one space dimension and the time dimension. The figure shows a graph with time plotted on the vertical axis and the one space dimension plotted on the horizontal axis. An *event* is something that occurs at a particular time and a particular point in space. ("Julius X. wrecks his car in Lemitar, NM on 21 June at 6:17 PM.") A *world line* is a plot of the position of some object as a function of time (more properly, the time of the object as a function of position) on a spacetime diagram. Thus, a world line is really a line in spacetime, while an event is a point in spacetime. A horizontal line parallel to the position axis (x-axis) is a *line of simultaneity*; in Galilean Relativity all events on this line occur simultaneously for all observers. It will be seen that the line of simultaneity differs between Galilean and Special Relativity; in Special Relativity the line of simultaneity depends on the state of motion of the observer.

In a spacetime diagram the slope of a world line has a special meaning. Notice that a vertical world line means that the object it represents does not move -- the velocity is zero. If the object moves to the right, then the world line tilts to the right, and the faster it moves, the more the world line tilts. Quantitatively, we say that

$$velocity = \frac{1}{slope\ of\ world\ line}. \quad (5.1)$$

Notice that this works for negative slopes and velocities as well as positive ones. If the object changes its velocity with time, then the world line is curved, and the instantaneous velocity at any time is the inverse of the slope of the tangent to the world line at that time.

The hardest thing to realize about spacetime diagrams is that they represent the past, present, and future all in one diagram. Thus, spacetime diagrams don't change with time -- the evolution of physical systems is represented by looking at successive horizontal slices in the diagram at successive times. Spacetime diagrams represent the evolution of events, but they don't evolve themselves.

The lightcone

Things that move at the speed of light in our four dimensional universe have surprising properties. If something travels at the speed of light along the x-axis and covers x meters from the origin in t seconds the space-time interval of its path is zero.

$s^2 = x^2 - (ct)^2$

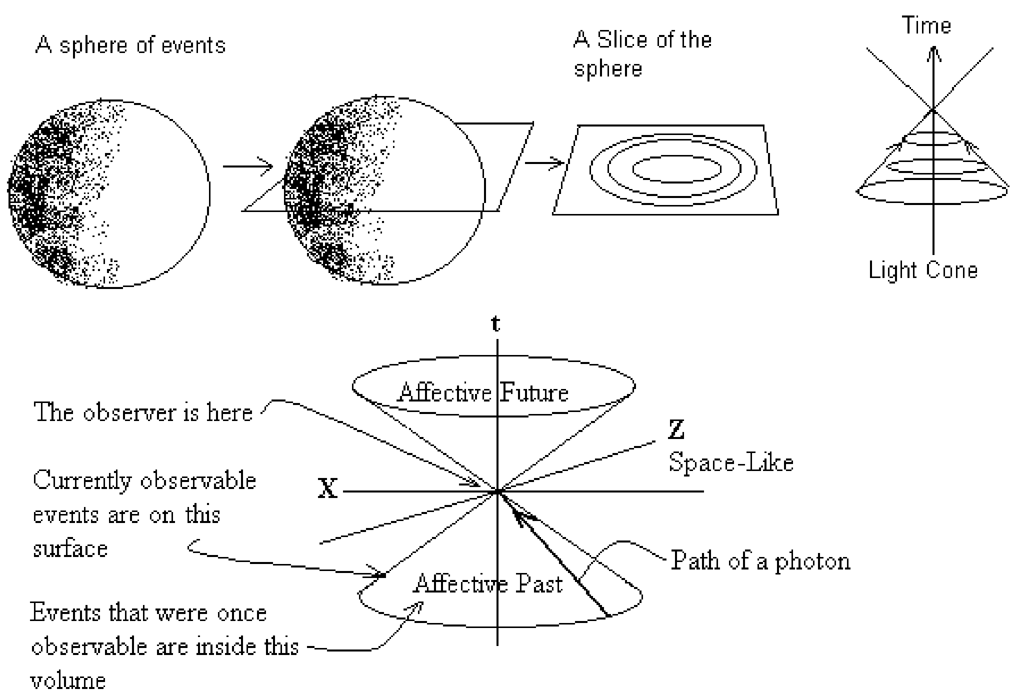
but $x = ct$ so:

$s^2 = (ct)^2 - (ct)^2 = 0$

Extending this result to the general case, if something travels at the speed of light in any direction into or out from the origin it has a space-time interval of 0:

$0 = x^2 + y^2 + z^2 - (ct)^2$

This equation is known as the Minkowski Light Cone Equation. If light were travelling towards the origin then the Light Cone Equation would describe the position and time of emission of all those photons that could be at the origin at a particular instant. If light were travelling away from the origin the equation would describe the position of the photons emitted at a particular instant at any future time 't'.



A Slice of a Sphere of Events Represented by a Light Cone (Solution of $0 = x^2 + z^2 - (ct)^2$)

At the superficial level the light cone is easy to interpret. It's backward surface represents the path of light rays that strike a point observer at an instant and it's forward surface represents the possible paths of rays emitted from the point observer at an instant (assuming the conditions appropriate to a special relativistic treatment prevail). Things that travel along the surface of the light cone are said to be **light- like** and the path taken by such things is known as a **null geodesic**.

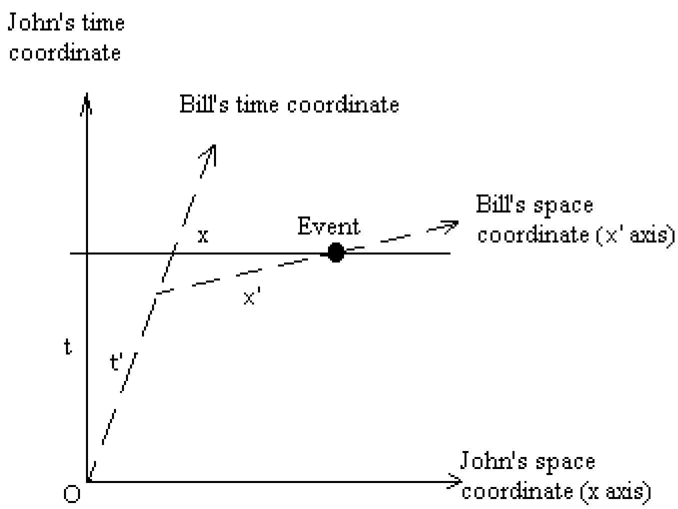
Events that lie outside the cones are said to be **space-like** or, better still **space separated** because their space time interval from the observer has the same sign as space (positive according to the convention used here). Events that lie within the cones are said to be **time-like** or **time separated** because their space-time interval has the same sign as time.

However, there is more to the light cone than the propagation of light. If the added assumption is made that the speed of light is the maximum possible velocity then events that are space separated cannot affect the observer directly. Events within the backward cone can have affected the observer so the backward cone is known as the "affective past" and the observer can affect events in the forward cone hence the forward cone is known as the "affective future".

The assumption that the speed of light is the maximum velocity for all communications is neither inherent in nor required by four dimensional geometry although the speed of light is indeed the maximum velocity for objects if the principle of **causality** is to be preserved by physical theories (ie: that causes precede effects).

The Lorentz transformation equations

The discussion so far has involved the comparison of interval measurements (time intervals and space intervals) between two observers. The observers might also want to compare more general sorts of measurement such as the time and position of a single event that is recorded by both of them. The equations that describe how each observer describes the other's recordings in this circumstance are known as the Lorentz Transformation Equations. (Note that the symbols below signify coordinates.)



The Lorentz Transformation: How observers with two different coordinate systems view the same event. In this case John and Bill are in space vehicles travelling past each other at 'O' where they synchronise clocks.

The objective is to calculate what the other observer reports as the time and position of the event

The table below shows the Lorentz Transformation Equations.

$x' = \frac{x - vt}{\sqrt{(1 - v^2/c^2)}}$	$x = \frac{x' + vt'}{\sqrt{(1 - v^2/c^2)}}$
$y' = y$	$y = y'$
$z' = z$	$z = z'$
$t' = \frac{t - (v/c^2)x}{\sqrt{(1 - v^2/c^2)}}$	$t = \frac{t' + (v/c^2)x'}{\sqrt{(1 - v^2/c^2)}}$

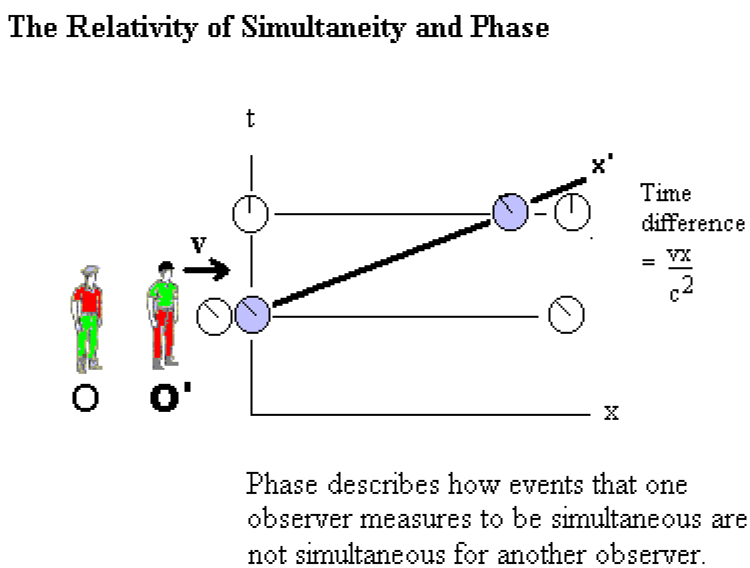
See mathematical derivation of Lorentz transformation.

Notice how the phase ($(v/c^2)x$) is important and how these formulae for absolute time and position of a joint event differ from the formulae for intervals.

Simultaneity, time dilation and length contraction

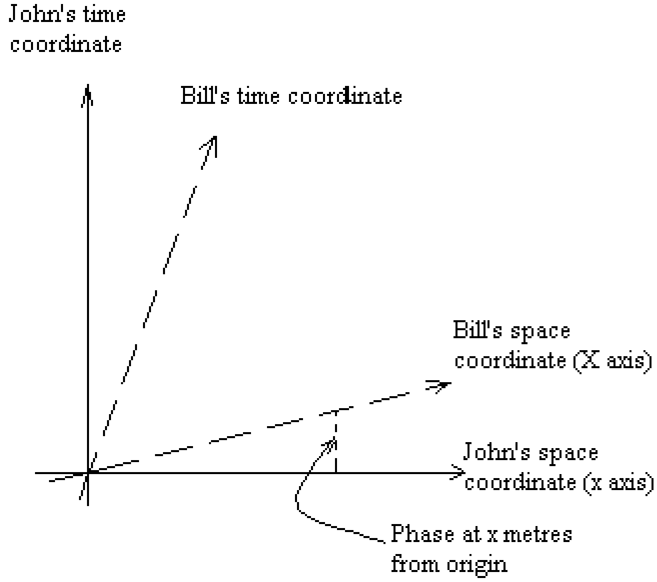
More about the relativity of simultaneity and the Andromeda paradox

If two observers who are moving relative to each other synchronise their clocks in their own frames of reference they discover that the clocks do not agree between the reference frames. This is illustrated below:



The effect of the relativity of simultaneity, or "phase", is for each observer to consider that a different set of events is simultaneous. Phase means that observers who are moving relative to each other have different sets of things that are simultaneous, or in their "present moment".

Figure 5: How Bill's coordinates appear to John at the instant Bill passes him.



How John views Bill's coordinate system

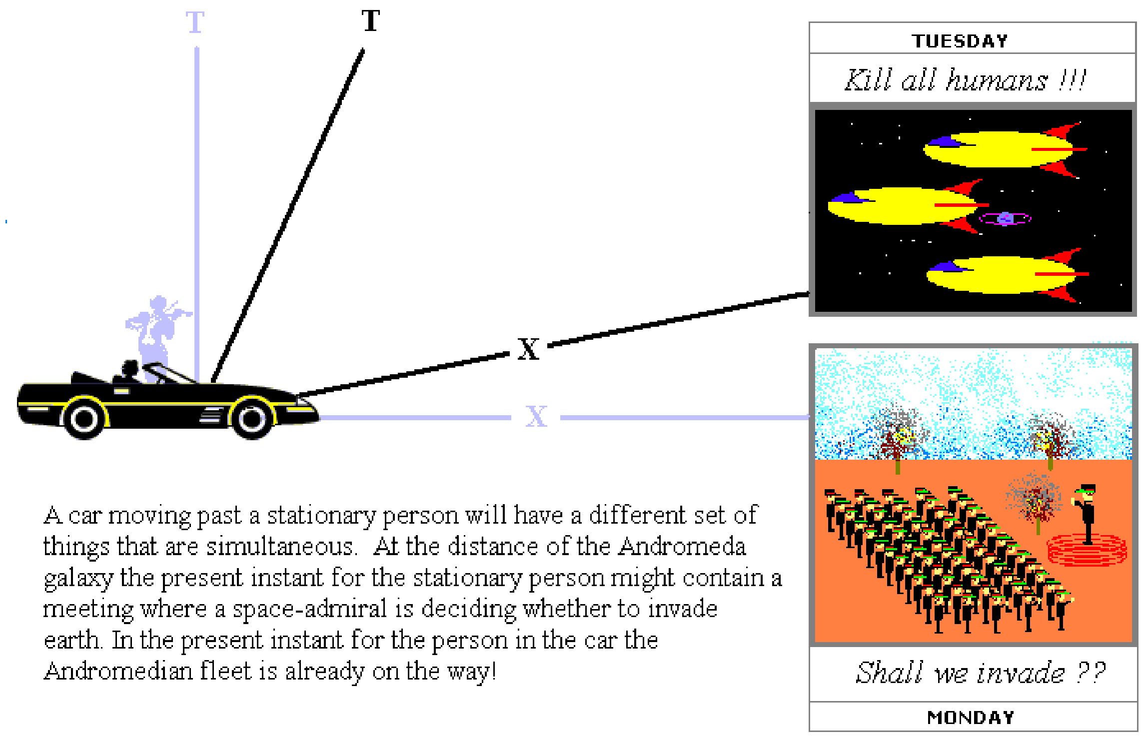
The amount by which the clocks differ between two observers depends upon the distance of the clock from the observer ($t = xv / c^2$). Notice that if both observers are part of inertial frames of reference with clocks that are synchronised at every point in space then the phase difference can be obtained by simply reading the difference between the clocks at the distant point and clocks at the origin. This difference will have the same value for both observers.

Relativistic phase differences have the startling consequence that at distances as large as our separation from nearby galaxies an observer who is driving on the earth can have a radically different set of events in her "present moment" from another person who is standing on the earth. The classic example of this effect of phase is the "Andromeda Paradox", also known as the "Rietdijk-Putnam-Penrose" argument. Penrose described the argument:

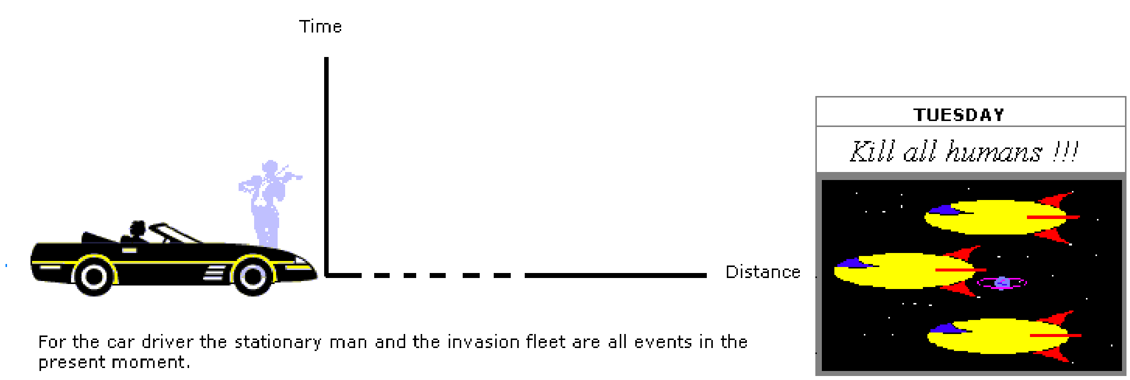
"Two people pass each other on the street; and according to one of the two people, an Andromedean space fleet has already set off on its journey, while to the other, the decision as to whether or not the journey will actually take place has not yet been made. How can there still be some uncertainty as to the outcome of that decision? If to either person the decision has already been made, then surely there cannot be any uncertainty. The launching of the space fleet is an inevitability." (Penrose 1989).

The argument is illustrated below:

The Andromeda Paradox



This "paradox" has generated considerable philosophical debate on the nature of time and free-will. A result of the relativity of simultaneity is that if the car driver launches a space rocket towards the Andromeda galaxy it might have a several days head start compared with a space rocket launched from the ground. This is because the "present moment" for the moving car driver is progressively advanced with distance compared with the present moment on the ground. The present moment for the car driver is shown in the illustration below:

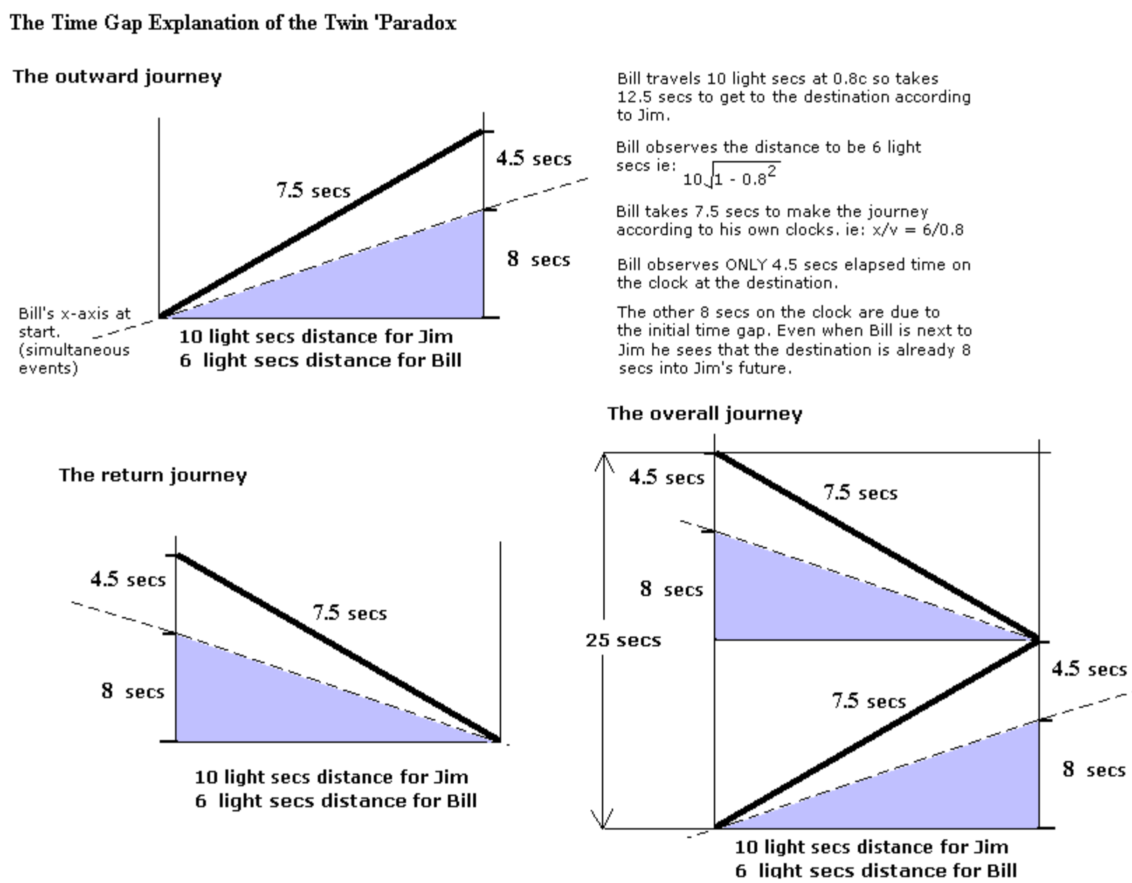


The twin paradox

The "Andromeda paradox" is, in part, the origin of the "twin paradox". In the twin paradox there are twins, Bill and Jim. Jim is on Earth. Bill flies past Jim in a spaceship, goes to a distant point, turns round and flies back again. It is found that Bill records fewer clock ticks over the whole journey than Jim records on earth. Why?

Suppose Jim has synchronised clocks on Earth and on the distant point. As Bill flies past Jim he synchronises his clock with Jim's clock. When he does this he observes the clocks on the distant point and immediately detects that they are not synchronised with his or Jim's clocks. To Bill it appears that Jim has synchronised his clocks incorrectly. There is a time difference, or "gap", between his clocks and those at the distant point even when he passes Jim. Bill flies to the distant point and discovers that the clock there is reading a later time than his own clock. He turns round to fly back to Earth and observes that the clocks on Earth seem to have jumped forward, yet another "time gap" appears. When Bill gets back to Earth the time gaps and time dilations mean that people on Earth have recorded more clock ticks that he did.

For ease of calculation suppose that Bill is moving at a truly astonishing velocity of 0.8c in the direction of a distant point that is 10 light seconds away (about 3 million kilometres). The illustration below shows Jim and Bill's observations:



The time dilation in the twin paradox is symmetrical. Jim observes Bill's clocks go slow, reading only 15 secs change on Bill's clocks when his own read that 25 secs have passed. Bill observes only 9 secs elapse on Jim's clocks when his own show 15 secs have passed. However Bill also observes Jim's clocks jump by 16 secs as a result of phase making a total 25 secs for Jim's clocks.

From Bill's viewpoint there is both a time dilation and a phase effect. It is the added factor of "phase" that explains why, although the time dilation occurs for both observers, Bill observes the same readings on Jim's clocks over the whole journey as does Jim.

To summarise the mathematics of the twin paradox using the example:

Jim observes the distance as 10 light seconds and the distant point is in his frame of reference. According to Jim it takes Bill the following time to make the journey:

Time taken = distance / velocity therefore according to Jim:

t = 10 / 0.8 = 12.5 seconds

Again according to Jim, time dilation should affect the observed time on Bill's clocks:

$T = t * \sqrt{1 - v^2/c^2}$ so:

$T = 12.5 * \sqrt{1 - 0.8^2} = 7.5$ seconds

So for Jim the round trip takes 25 secs and Bill's clock reads 15 secs.

Bill measures the distance as:

$X = x * \sqrt{1 - v^2/c^2} = 10 * \sqrt{1 - 0.8^2} = 6$ light seconds.

For Bill it takes X / v = 6 / 0.8 = 7.5 seconds.

Bill observes Jim's clocks to appear to run slow as a result of time dilation:

$t' = T * \sqrt{1 - v^2/c^2}$ so:

$t' = 7.5 * \sqrt{1 - 0.8^2} = 4.5$ seconds

But there is also a time gap of vx / c^2 = 8 seconds.

So for Bill, Jim's clocks register 12.5 secs have passed from the start to the distant point. This is composed of 4.5 secs elapsing on Jim's clocks plus an 8 sec time gap from the start of the journey. Bill sees 25 secs total time recorded on Jim's clocks over the whole journey, this is the same time as Jim observes on his own clocks.

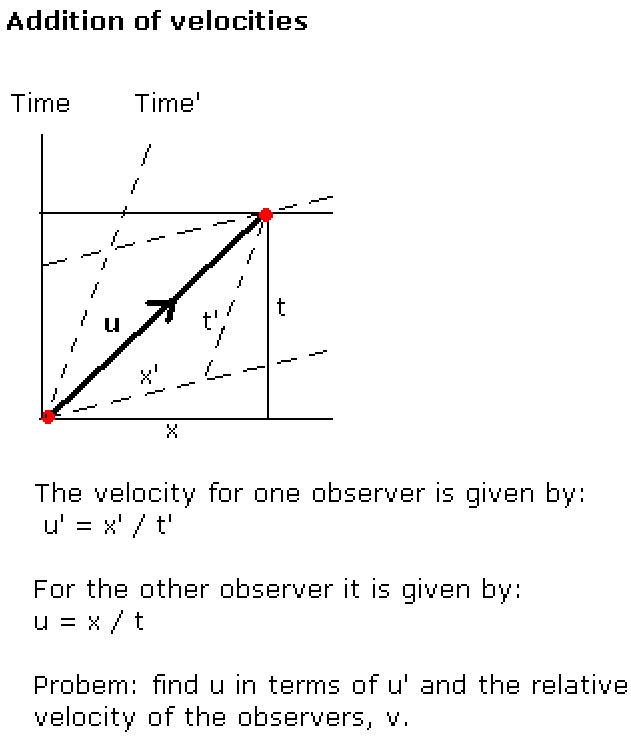
It is sometimes dubiously asserted that the twin paradox is about the clocks on the twin that leaves earth being slower than those on the twin that stays at home, it is then argued that biological processes contain clocks therefore the twin that travelled away ages less. A more accurate explanation is that when we travel we travel in time as well as space.

The Pole-barn paradox

this is a stub and requires completion

Addition of velocities

How can two observers, moving at v km/sec relative to each other, compare their observations of the velocity of a third object?



Suppose one of the observers measures the velocity of the object as u' where:

$$u' = \frac{x'}{t'}$$

The coordinates x' and t' are given by the Lorentz transformations:

$$x' = \frac{x - vt}{\sqrt{(1 - v^2/c^2)}}$$

and

$$t' = \frac{t - (v/c^2)x}{\sqrt{(1 - v^2/c^2)}}$$

but

$x' = u't'$

so:

$$\frac{x - vt}{\sqrt{(1 - v^2/c^2)}} = u' \frac{t - (v/c^2)x}{\sqrt{(1 - v^2/c^2)}}$$

and hence:

$$x - vt = u'(t - vx / c^2)$$

Notice the role of the phase term vx / c^2. The equation can be rearranged as:

$$x = \frac{(u' + v)}{(1 + u'v/c^2)}t$$

given that $x = ut$:

$$u = \frac{(u' + v)}{(1 + u'v/c^2)}$$

This is known as the **relativistic velocity addition theorem**, it applies to velocities parallel to the direction of mutual motion.

The existence of time dilation means that even when objects are moving perpendicular to the direction of motion there is a discrepancy between the velocities reported for an object by observers who are moving relative to each other. If there is any component of velocity in the x direction (u_x, u'_x) then the phase affects time measurement and hence the velocities perpendicular to the x-axis. The table below summarises the relativistic addition of velocities in the various directions in space.

$u'_x = \frac{(u_x - v)}{(1 - u_x v/c^2)}$	$u_x = \frac{(u'_x + v)}{(1 + u'_x v/c^2)}$
$u'_y = \frac{u_y\sqrt{1 - v^2/c^2}}{(1 - u_x v/c^2)}$	$u_y = \frac{u'_y\sqrt{1 - v^2/c^2}}{(1 + u'_x v/c^2)}$
$u'_z = \frac{u_z\sqrt{1 - v^2/c^2}}{(1 - u_x v/c^2)}$	$u_z = \frac{u'_z\sqrt{1 - v^2/c^2}}{(1 + u'_x v/c^2)}$

Notice that for an observer in another reference frame the sum of two velocities (u and v) can never exceed the speed of light.

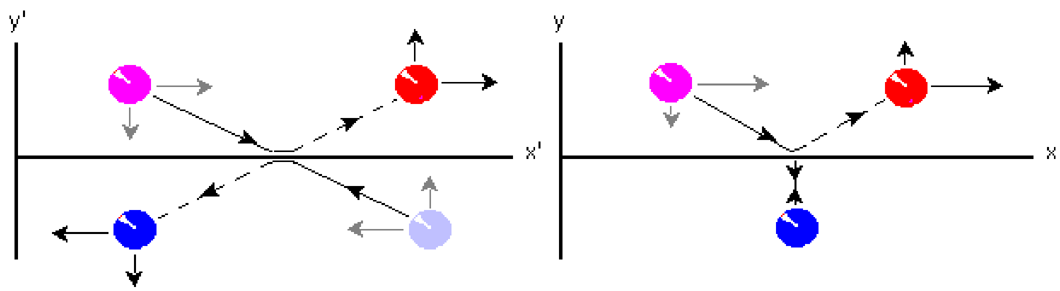
Dynamics

Introduction

The way that the velocity of a particle can differ between observers who are moving relative to each other means that momentum needs to be redefined as a result of relativity theory.

The illustration below shows a typical collision of two particles. In the right hand frame the collision is observed from the viewpoint of someone moving at the same velocity as one of the particles, in the left hand frame it is observed by someone moving at a velocity that is intermediate between those of the particles.

Relativistic particle collisions



View of observer who is travelling at a velocity intermediate between that of the balls. (Red = +v Blue = -v Observer =0 m/s)

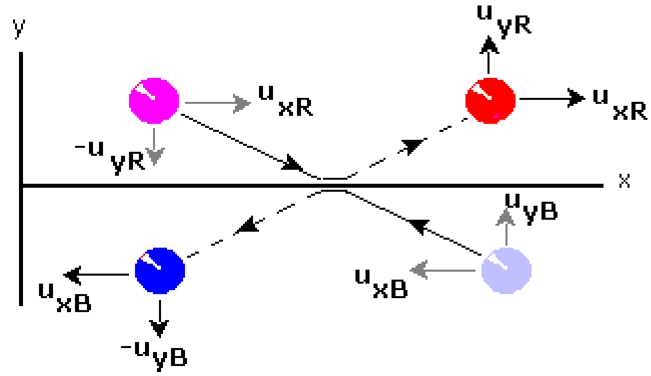
View of observer travelling at same velocity as the blue ball.

If momentum is redefined then all the variables such as force (rate of change of momentum), energy etc. will become redefined and relativity will lead to an entirely new physics. The new physics has an effect at the ordinary level of experience through the relation $E = mc^2$ whereby it is the tiny changes in relativistic mass that are expressed as everyday kinetic energy so that the whole of physics is related to "relativistic" reasoning rather than Newton's empirical ideas.

Momentum

In physics momentum is conserved within a closed system, the **law of conservation of momentum** applies. Consider the special case of identical particles colliding symmetrically as illustrated below:

A symmetrical Newtonian collision



The momentum change by the red ball is:

$$2\mathbf{u}_{\mathbf{yR}}m$$

The momentum change by the blue ball is:

$$2\mathbf{u}_{\mathbf{yB}}m$$

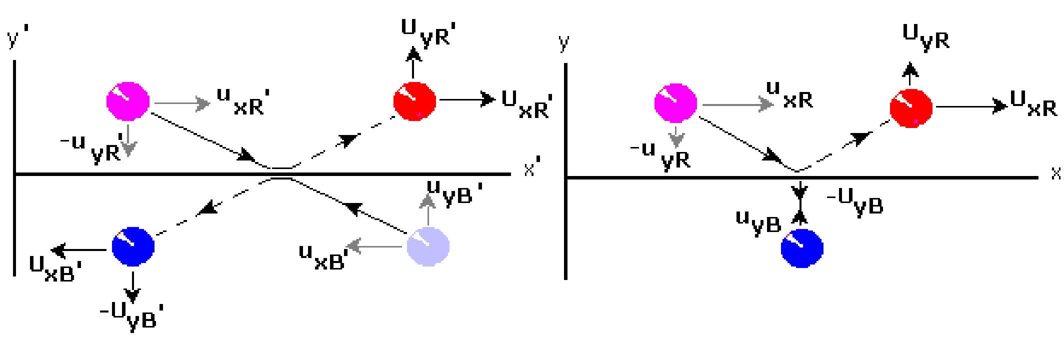
The situation is symmetrical so the **Newtonian** conservation of momentum law is demonstrated:

$$2m\mathbf{u}_{\mathbf{yR}} = 2m\mathbf{u}_{\mathbf{yB}}$$

Notice that this result depends upon the y components of the velocities being equal ie: $\mathbf{u}_{\mathbf{yR}} = \mathbf{u}_{\mathbf{yB}}$.

The relativistic case is rather different. The collision is illustrated below, the left hand frame shows the collision as it appears for one observer and the right hand frame shows **exactly the same collision** as it appears for another observer moving at the same velocity as the blue ball:

Relativistic particle collisions



View of observer who is travelling at a velocity intermediate between that of the balls. (Red = +v Blue = -v Observer =0 m/s)

View of observer travelling at same velocity as the blue ball.

The uppercase letters (U) represent velocities after the collision.

The configuration shown above has been simplified because one frame contains a stationary blue ball (ie: u_{xB} = 0) and the velocities are chosen so that the vertical velocity of the red ball is exactly reversed after the collision ie: $u'_{yR} = -u'_{yB}$. Both frames show exactly the same event, it is only the observers who differ between frames. The relativistic velocity transformations between frames is:

$$u'_{yR} = \frac{u_{yR}\sqrt{1 - v^2/c^2}}{1 - u_{xR}v/c^2}$$

$u'_{yB} = u_{yB}\sqrt{1 - v^2/c^2}$ given that $u_{xB} = 0$.

Suppose that the y components are equal in one frame, in Newtonian physics they will also be equal in the other frame. However, in relativity, if the y components are equal in one frame they are **not** necessarily equal in the other frame. For instance if $u'_{yR} = u'_{yB}$ then:

$$u_{yB} = \frac{u_{yR}}{1 - u_{xR}v/c^2}$$

So if $u'_{yR} = u'_{yB}$ then in this case $u_{yR} \neq u_{yB}$.

If the mass were constant between collisions and between frames then although $2m\mathbf{u'_{yR}} = 2m\mathbf{u'_{yB}}$ it is found that:

$$2m\mathbf{u_{yR}} \neq 2m\mathbf{u_{yB}}$$

So momentum would not appear to be conserved between frames if the mass is constant. Notice that the discrepancy is very small if u_{xR} and v are small. However, the principle of relativity states that **the laws of physics are the same in all inertial systems**, so to preserve this principle there must be something happening to the mass as observed between frames.

The velocities in the y direction are related by the following equation when the observer is travelling at the same velocity as the blue ball ie: when $u_{xB} = 0$:

$$u_{yB} = \frac{u_{yR}}{1 - u_{xR}v/c^2}$$

If we write m_B for the mass of the blue ball) and m_R for the mass of the red ball as observed from the frame of the blue ball then, if the principle of relativity applies:

$$2m_R u_{yR} = 2m_B u_{yB}$$

So:

$$m_R = m_B \frac{u_{yB}}{u_{yR}}$$

But:

$$u_{yB} = \frac{u_{yR}}{1 - u_{xR}v/c^2}$$

Therefore:

$$m_R = \frac{m_B}{1 - u_{xR}v/c^2}$$

This means that, if the principle of relativity is to apply then the mass must change by the amount shown in the equation above for the conservation of momentum law to be true.

The reference frame was chosen so that $u'_{yR} = -u'_{yB}$ and hence $u'_{xR} = v$. This allows *v* to be determined in terms of u_{xR}:

$$u'_{xR} = \frac{u_{xR} - v}{1 - u_{xR}v/c^2} = v$$

and hence:

$$v = c^2/u_{xR}(1 - \sqrt{1 - u_{xR}^2/c^2}$$

So substituting for *v* in $m_R = \frac{m_B}{1 - u_{xR}v/c^2}$:

$$m_R = \frac{m_B}{\sqrt{1 - u_{xR}^2/c^2}}$$

The blue ball is at rest so its mass is its **rest mass**, and is given the symbol m_0. As the balls were identical at the start of the boost the mass of the red ball is the mass that a blue ball would have if it were in motion relative to an observer; this mass is known as the **relativistic mass** symbolised by *m*. The discussion given above was related to the relative motions of the blue and red balls, as a result u_{xR} corresponds to the **speed** of the moving ball relative to an observer who is stationary with respect to the blue ball. These considerations mean that the relativistic mass is given by:

$$\mathbf{m} = \frac{\mathbf{m_0}}{\sqrt{\mathbf{1 - u^2/c^2}}}$$

The relativistic momentum is given by the product of the relativistic mass and the velocity $\mathbf{p} = m\mathbf{u}$.

The overall expression for momentum in terms of rest mass is:

$$\mathbf{p} = \frac{m_0\mathbf{u}}{\sqrt{1 - u^2/c^2}}$$

and the components of the momentum are:

$$p_x = \frac{m_0 u_x}{\sqrt{1 - u^2/c^2}}$$

$$p_y = \frac{m_0 u_y}{\sqrt{1 - u^2/c^2}}$$

$$p_z = \frac{m_0 u_z}{\sqrt{1 - u^2/c^2}}$$

So the components of the momentum depend upon the appropriate velocity component and the speed.

Force

In Newtonian mechanics force is the rate of change of momentum ($d\mathbf{p}/dt$. If the relativistic momentum is used:

$$\frac{d\mathbf{p}}{dt} = \frac{d(m\mathbf{u})}{dt}$$

By Leibniz's law where $d(xy) = xdy + ydx$:

$$\frac{d\mathbf{p}}{dt} = m\frac{du}{dt} + u\frac{dm}{dt}$$

This shows that part of the force is used to increase the velocity and part is used to increase the relativistic mass. Relativistic force is different from Newtonian force ($\mathbf{f} = md\mathbf{u}/dt$).

Energy

Energy is defined as the work done in moving a body from one place to another. Energy is given from:

$$dE = \mathbf{F}d\mathbf{x}$$

so, over the whole path:

$$E = \int_0^x \mathbf{F}d\mathbf{x}$$

Kinetic energy (K) is the energy used to move a body from a velocity of 0 to a velocity *u*. So:

$$K = \int_{u=0}^{u=u} F dx$$

Using the relativistic force:

$$K = \int_{u=0}^{u=u} \frac{d(mu)}{dt} dx$$

So:

$$K = \int_{u=0}^{u=u} d(mu) \frac{dx}{dt}$$

substituting for *d*(*mu*) and using *dx* / *dt* = *u*:

$$K = \int_{u=0}^{u=u} (mdu + udm)u$$

Which gives:

$$K = \int_{u=0}^{u=u} (mudu + u^2 dm)$$

The relativistic mass is given by:

$$\mathbf{m} = \frac{\mathbf{m_0}}{\sqrt{\mathbf{1 - u^2/c^2}}}$$

Which can be expanded as:

$$m^2c^2 - m^2u^2 = m_0^2c^2$$

Differentiating:

$$2mc^2dm - m^2 2udu - u^2 2mdm = 0$$

So, rearranging:

$$mudu + u^2dm = c^2dm$$

In which case:

$$K = \int_{u=0}^{u=u} (mudu + u^2 dm)$$

is simplified to:

$$K = \int_{u=0}^{u=u} c^2 dm$$

But the mass goes from m_0 to *m* so:

$$K = c^2 \int_{m=m_0}^{m=m} dm)$$

and hence:

$$\mathbf{K = mc^2 - m_0c^2}$$

The amount mc^2 is known as the **total energy** of the particle. The amount m_0c^2 is known as the **rest energy** of the particle. If the total energy of the particle is given the symbol E:

$$\mathbf{E = m_0c^2 + K}$$

So it can be seen that m_0c^2 is the energy of a mass that is stationary. This energy is known as **mass energy** and is the origin of the famous formula $E = mc^2$ that is iconic of the nuclear age.

The Newtonian approximation for kinetic energy can be derived by substituting the rest mass for the relativistic mass ie:

$$\mathbf{m = \frac{m_0}{\sqrt{1 - u^2/c^2}}}$$

and:

$$\mathbf{K = mc^2 - m_0c^2}$$

So:

$$K = \frac{m_0c^2}{\sqrt{1 - u^2/c^2}} - m_0c^2$$

ie:

$$K = m_0c^2((1 - u^2/c^2)^{-\frac{1}{2}} - 1)$$

The binomial theorem can be used to expand $(1 - u^2/c^2)^{-\frac{1}{2}}$:

The binomial theorem is:

$$(a + x)^n = a^n + na^{n-1}x + \frac{n(n-1)}{2!}a^{n-2}x^2 \ldots.$$

So expanding $(1 - u^2/c^2)^{-\frac{1}{2}}$:

$$K = \frac{1}{2}m_0u^2 + \frac{3m_0u^4}{8c^2} + \frac{5m_0v^6}{16c^4} + \ldots$$

So if u is much less than c:

$$K = \frac{1}{2} m_0 u^2$$

Which is the Newtonian approximation for low velocities.

Aether

Introduction

Many students confuse Relativity Theory with a theory about the propagation of light. According to modern Relativity Theory the constancy of the speed of light is a consequence of the geometry of spacetime rather than something specifically due to the properties of photons; but the statement "the speed of light is constant" often distracts the student into a consideration of light propagation. This confusion is amplified by the importance assigned to interferometry experiments, such as the Michelson-Morley experiment, in most textbooks on Relativity Theory.

The history of theories of the propagation of light is an interesting topic in physics and was indeed important in the early days of Relativity Theory. In the seventeenth century two competing theories of light propagation were developed. Christiaan Huygens published a **wave theory of light** which was based on **Huygen's principle** whereby every point in a wavelike disturbance can give rise to further disturbances that spread out spherically. In contrast Newton considered that the propagation of light of light was due to the passage of small particles or "corpuscles" from the source to the illuminated object. His theory is known as the **corpuscular theory of light**. Newton's theory was widely accepted until the nineteenth century.

In the early nineteenth century Thomas Young performed his **Young's slits** experiment and the interference pattern that occurred was explained in terms of diffraction due to the wave nature of light. The wave theory was accepted generally until the twentieth century when quantum theory confirmed that light had a corpuscular nature and that Huygen's principle could not be applied.

The idea of light as a disturbance of some medium, or **aether**, that permeates the universe was problematical from its inception. The first problem that arose was that the speed of light did not change with the velocity of the observer. If light were indeed a disturbance of some stationary medium then as the earth moves through the medium towards a light source the speed of light should appear to increase. It was found however that the speed of light did not change as expected. Each experiment on the velocity of light required corrections to existing theory and led to a variety of subsidiary theories such as the "aether drag hypothesis". Ultimately it was experiments that were designed to investigate the properties of the aether that provided the first experimental evidence for Relativity Theory.

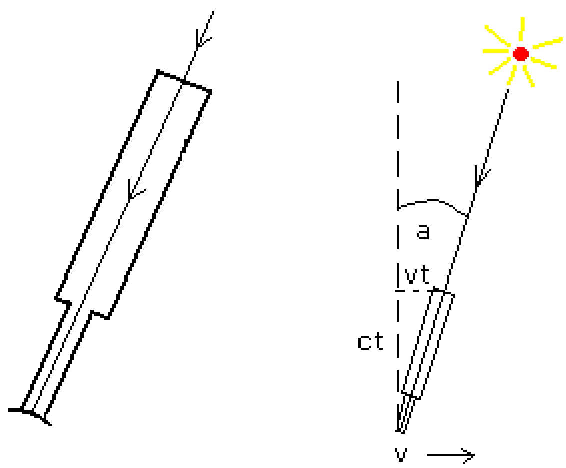
The aether drag hypothesis

The **aether drag hypothesis** was an early attempt to explain the way experiments such as Arago's experiment showed that the speed of light is constant. The aether drag hypothesis is now considered to be incorrect by mainstream science.

According to the aether drag hypothesis light propagates in a special medium, the aether, that remains attached to things as they move. If this is the case then, no matter how fast the earth moves around the sun or rotates on its axis, light on the surface of the earth would travel at a constant velocity.

The primary reason the aether drag hypothesis is considered invalid is because of the occurrence of stellar aberration. In stellar aberration the position of a star when viewed with a telescope swings each side of a central position by about 20.5 seconds of arc every six months. This amount of swing is the amount expected when considering the speed of earth's travel in its orbit. In 1871 George Biddell Airy demonstrated that stellar aberration occurs even when a telesope is filled with water. It seems that if the aether drag hypothesis were true then stellar aberration would not occur because the light would be travelling in the aether which would be moving along with the telescope.

Stellar Aberration



tan(a) = vt/ct

If a telescope is travelling at high speed only light that is arranged at a particular angle can avoid hitting the walls of the telescope tube

If you visualize a bucket on a train about to enter a tunnel and a drop of water drips from the tunnel entrance into the bucket at the very center, the drop will not hit the center at the bottom of the bucket. The bucket is the tube of a telescope, the drop is a photon and the train is the earth. If aether is dragged then the droplet would be traveling with the train when it is dropped and would hit the center of bucket at the bottom.

The amount of stellar aberration, α is given by:

$$tan(\alpha) = v\delta t \,/\, c\delta t$$

So:

$$tan(\alpha) = v \,/\, c$$

The speed at which the earth goes round the sun, v = 30 km/s, and the speed of light is c = 300,000,000 m/s which gives α = 20.5 seconds of arc every six months. This amount of aberration is observed and this contradicts the aether drag hypothesis.

In 1818 Fresnel introduced a modification to the aether drag hypothesis that only applies to the interface between media. This was accepted during much of the nineteenth century but has now been replaced by special theory of relativity (see below).

The aether drag hypothesis is historically important because it was one of the reasons why Newton's corpuscular theory of light was replaced by the wave theory and it is used in early explanations of light propagation without relativity theory. It originated as a result of early attempts to measure the speed of light.

In 1810 François Arago realised that variations in the refractive index of a substance predicted by the corpuscular theory would provide a useful method for measuring the velocity of light. These predictions arose because the refractive index of a substance such as glass depends on the ratio of the velocities of light in air and in the glass. Arago attempted to measure the extent to which corpuscles of light would be refracted by a glass prism at the front of a telescope. He expected that there would be a range of different angles of refraction due to the variety of different velocities of the stars and the motion of the earth at different times of the day and year. Contrary to this expectation he found that that there was no difference in refraction between stars, between times of day or between seasons. All Arago observed was ordinary stellar aberration.

In 1818 Augustin Jean Fresnel examined Arago's results using a wave theory of light. He realised that even if light were transmitted as waves the refractive index of the glass-air interface should have varied as the glass moved through the aether to strike the incoming waves at different velocities when the earth rotated and the seasons changed.

Fresnel proposed that the glass prism would carry some of the aether along with it so that "..the aether is in excess inside the prism". He realised that the velocity of propagation of waves depends on the density of the medium so proposed that the velocity of light in the prism would need to be adjusted by an amount of 'drag'.

The velocity of light v_n in the glass without any adjustment is given by:

$$v_n = c / n$$

The drag adjustment v_d is given by:

$$v_d = v(1 - \frac{\rho_e}{\rho_g})$$

Where ρ_e is the aether density in the environment, ρ_g is the aether density in the glass and v is the velocity of the prism with respect to the aether.

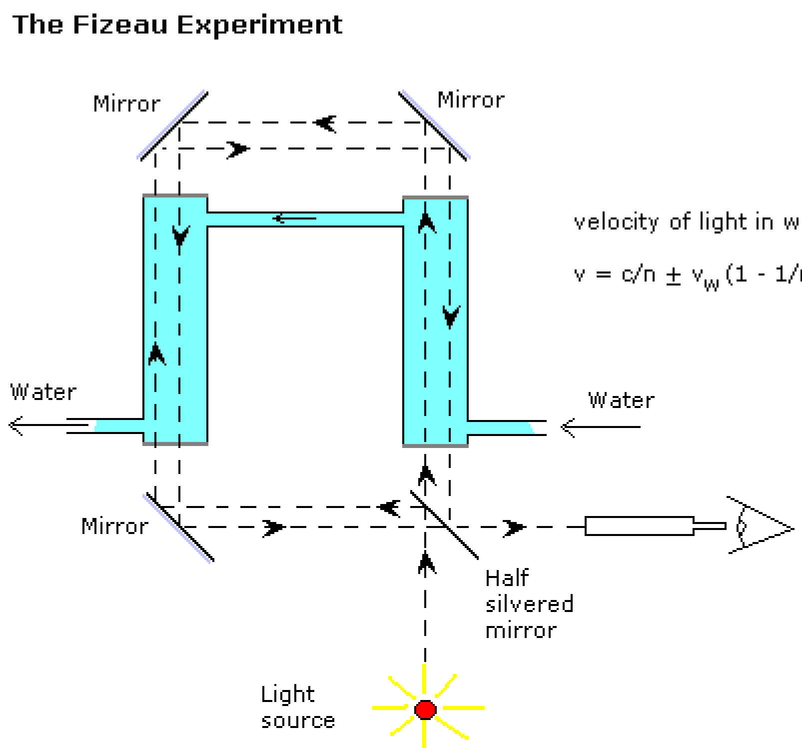
The factor $(1 - \frac{\rho_e}{\rho_g})$ can be written as $(1 - \frac{1}{n^2})$ because the refractive index, n, would be dependent on the density of the aether. This is known as the **Fresnel drag coefficient**.

The velocity of light in the glass is then given by:

$$V = \frac{c}{n} + v(1 - \frac{1}{n^2})$$

This correction was successful in explaining the null result of Arago's experiment. It introduces the concept of a largely stationary aether that is dragged by substances such as glass but not by air. Its success favoured the wave theory of light over the previous corpuscular theory.

The Fresnel drag coefficient was confirmed by an interferometer experiment performed by Fizeau. Water was passed at high speed along two glass tubes that formed the optical paths of the interferometer and it was found that the fringe shifts were as predicted by the drag coefficient.



The special theory of relativity predicts the result of the Fizeau experiment from the velocity addition theorem without any need for an aether.

If V is the velocity of light relative to the Fizeau apparatus and U is the velocity of light relative to the water and v is the velocity of the water:

$$U = \frac{c}{n}$$

$$V = \frac{c/n + v}{1 + v/nc}$$

which, if v/c is small can be expanded using the binomial expansion to become:

$$V = \frac{c}{n} + v(1 - \frac{1}{n^2})$$

This is identical to Fresnel's equation.

It may appear as if Fresnel's analysis can be substituted for the relativistic approach, however, more recent work has shown that Fresnel's assumptions should lead to different amount of aether drag for different frequencies of light and violate Snell's law (see Ferraro and Sforza (2005)).

The aether drag hypothesis was one of the arguments used in an attempt to explain the Michelson-Morley experiment before the widespread acceptance of the special theory of relativity.

The Fizeau experiment is consistent with relativity and approximately consistent with each individual body, such as prisms, lenses etc. dragging its own aether with it. This contradicts some modified versions of the aether drag hypothesis that argue that aether drag may happen on a global (or larger) scale and stellar aberration is merely transferred into the entrained "bubble" around the earth which then faithfully carries the modified angle of incidence directly to the observer.

References

- Rafael Ferraro and Daniel M Sforza 2005. Arago (1810): the first experimental result against the ether Eur. J. Phys. 26 195-204

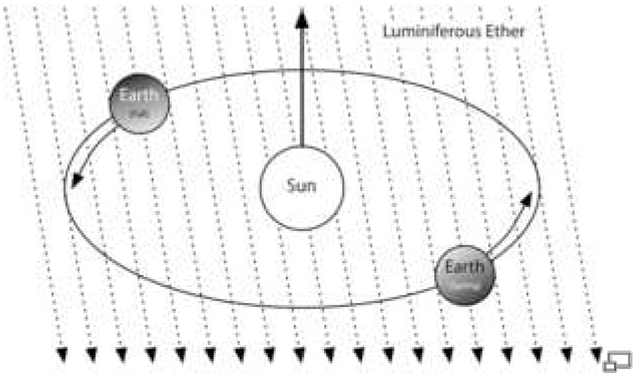
The Michelson-Morley experiment

(This article has been copied from Wikipedia)

The **Michelson-Morley experiment**, one of the most important and famous experiments in the history of physics, was performed in 1887 by Albert Michelson and Edward Morley at what is now Case Western Reserve University, and is considered to be the first strong evidence against the theory of a luminiferous aether.

Physics theories of the late 19th century postulated that, just as water waves must have a medium to move across (water), and audible sound waves require a medium to move through (air), so also light waves require a medium, the "luminiferous aether". The speed of light being so great, designing an experiment to detect the presence and properties of this aether took considerable thought.

Measuring aether



A depiction of the concept of the "aether wind".

Each year, the Earth travels a tremendous distance in its orbit around the sun, at a speed of around 30 km/second, over 100,000 km per hour. It was reasoned that the Earth would at all times be moving through the aether and producing a detectable "aether wind". At any given point on the Earth's surface, the magnitude and direction of the wind would vary with time of day and season. By analysing the effective wind at various different times, it should be possible to separate out components due to motion of the Earth relative to the Solar System from any due to the overall motion of that system.

The effect of the aether wind on light waves would be like the effect of wind on sound waves. Sound waves travel at a constant speed relative to the medium that they are travelling through (this varies depending on the pressure, temperature etc (see sound), but is typically around 340 m/s). So, if the speed of sound in our conditions is 340 m/s, when there is a 10 m/s wind relative to the ground, into the wind it will appear that sound is travelling at 330 m/s (340 - 10). Downwind, it will appear that sound is travelling at 350 m/s (340 + 10). Measuring the speed of sound compared to the ground in different directions will therefore enable us to calculate the speed of the air relative to the ground.

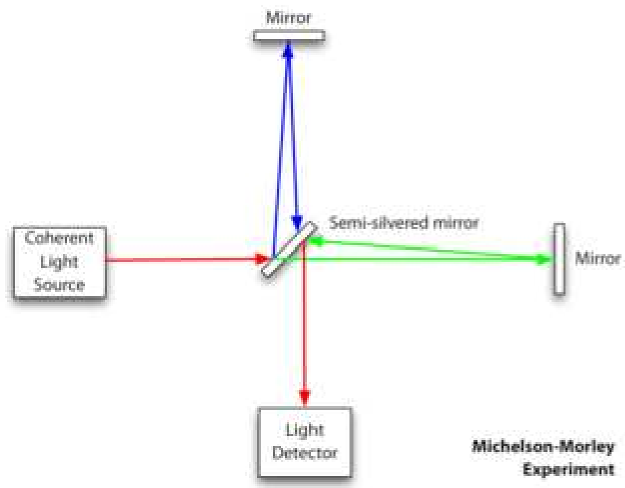
If the speed of the sound cannot be directly measured, an alternative method is to measure the time that the sound takes to bounce off of a reflector and return to the origin. This is done parallel to the wind and perpendicular (since the direction of the wind is unknown before hand, just determine the time for several different directions). The cumulative round trip effects of the wind in the two orientations slightly favors the sound travelling at right angles to it. Similarly, the effect of an aether wind on a beam of light would be for the beam to take slightly longer to travel round-trip in the direction parallel to the "wind" than to travel the same round-trip distance at right angles to it.

"Slightly" is key, in that, over a distance such as a few meters, the difference in time for the two round trips would be only about a millionth of a millionth of a second. At this point the only truly accurate measurements of the speed of light were those carried out by Albert Abraham Michelson, which had resulted in measurements accurate to a few meters per second. While a stunning achievement in its own right, this was certainly not nearly enough accuracy to be able to detect the aether.

The experiments

Michelson, though, had already seen a solution to this problem. His design, later known as an interferometer, sent a single source of white light through a half-silvered mirror that was used to split it into two beams travelling at right angles to one another. After leaving the splitter, the beams travelled out to the ends of long arms where they were reflected back into the middle on small mirrors. They then recombined on the far side of the splitter in an eyepiece, producing a pattern of constructive and destructive interference based on the length of the arms. Any slight change in the amount of time the beams spent in transit would then be observed as a shift in the positions of the interference fringes. If the aether were stationary relative to the sun, then the Earth's motion would produce a shift of about 0.04 fringes.

Michelson had made several measurements with an experimental device in 1881, in which he noticed that the expected shift of 0.04 was not seen, and a smaller shift of about 0.02 was. However his apparatus was a prototype, and had experimental errors far too large to say anything about the aether wind. For a measurement of the aether wind, a much more accurate and tightly controlled experiment would have to be carried out. The prototype was, however, successful in demonstrating that the basic method was feasible.



A Michelson interferometer

He then combined forces with Edward Morley and spent a considerable amount of time and money creating an improved version with more than enough accuracy to detect the drift. In their experiment the light was repeatedly reflected back and forth along the arms, increasing the path length to 11m. At this length the drift would be about .4 fringes. To make that easily detectable the apparatus was located in a closed room in the basement of a stone building, eliminating most thermal and vibrational effects. Vibrations were further reduced by building the apparatus on top of a huge block of marble, which was then floated in a pool of mercury. They calculated that effects of about 1/100th of a fringe would be detectable.

The mercury pool allowed the device to be turned, so that it could be rotated through the entire range of possible angles to the "aether wind". Even over a short period of time some sort of effect would be noticed simply by rotating the device, such that one arm rotated into the direction of the wind and the other away. Over longer periods day/night cycles or yearly cycles would also be easily measurable.

During each full rotation of the device, each arm would be parallel to the wind twice (facing into and away from the wind) and perpendicular to the wind twice. This effect would show readings in a sine wave formation with two peaks and two troughs. Additionally if the wind was only from the earth's orbit around the sun, the wind would fully change directions east/west during a 12 hour period. In this ideal conceptualization, the sine wave of day/night readings would be in opposite phase.

Because it was assumed that the motion of the solar system would cause an additional component to the wind, the yearly cycles would be detectable as an alteration of the maginitude of the wind. An example of this effect is a helicopter flying forward. While on the ground, a helicopter's blades would be measured as travelling around at 50 MPH at the tips. However, if the helicopter is travelling forward at 50 MPH, there are points at which the tips of the blades are travelling 0 MPH and 100 MPH with respect to the air they are travelling through. This increases the magnitude of the lift on one side and decreases it on the other just as it would increase and decrease the magnitude of an ether wind on a yearly basis.

The most famous failed experiment

Ironically, after all this thought and preparation, the experiment became what might be called the most famous failed experiment to date. Instead of providing insight into the properties of the aether, Michelson and Morley's 1887 article in the American Journal of Science reported the measurement to be as small as one-fortieth of the expected displacement but "since the displacement is proportional to the square of the velocity" they concluded that the measured velocity was approximately one-sixth of the expected velocity of the Earth's motion in orbit and "certainly less than one-fourth". Although this small "velocity" was measured, it was considered far too small to be used as evidence of aether, it was later said to be within the range of an experimental error that would allow the speed to actually be zero.

Although Michelson and Morley went on to different experiments after their first publication in 1887, both remained active in the field. Other versions of the experiment were carried out with increasing sophistication. Kennedy and Illingsworth both modified the mirrors to include a half-wave "step", eliminating the possibility of some sort of standing wave pattern within the apparatus. Illingsworth could detect changes on the order of 1/300th of a fringe, Kennedy up to 1/1500th. Miller later built a non-magnetic device to eliminate magnetostriction, while Michelson built one of non-expanding invar to eliminate any remaining thermal effects. Others from around the world increased accuracy, eliminated possible side effects, or both. All of these with the exception of Dayton Miller also returned what is considered a null result.

Morley was not convinced of his own results, and went on to conduct additional experiments with Dayton Miller. Miller worked on increasingly large experiments, culminating in one with a 32m (effective) arm length at an installation at the Mount Wilson observatory. To avoid the possibility of the aether wind being blocked by solid walls, he used a special shed with thin walls, mainly of canvas. He consistently measured a small positive effect that varied, as expected, with each rotation of the device, the sidereal day and on a yearly basis. The low magnitude of the results he attributed to aether entrainment (see below). His measurements amounted to only ~10 kps instead of the expected ~30 kps expected from the earth's orbital motion alone. He remained convinced this was due to *partial* entrainment, though he did not attempt a detailed explanation.

Though Kennedy later also carried out an experiment at Mount Wilson, finding 1/10 the drift measured by Miller, and no seasonal effects, Miller's findings were considered important at the time, and were discussed by Michelson, Hendrik Lorentz and others at a meeting reported in 1928 (ref below). There was general agreement that more experimentation was needed to check Miller's results. Lorentz recognised that the results, whatever their cause, did not quite tally with either his or Einstein's versions of special relativity. Einstein was not present at the meeting and felt the results could be dismissed as experimental error (see Shankland ref below).

Name	Year	Arm length (meters)	Fringe shift expected	Fringe shift measured	Experimental Resolution	Upper Limit on V_{aether}
Michelson	1881	1.2	0.04	0.02		
Michelson and Morley	1887	11.0	0.4	< 0.01		8 km/s
Morley and Morley	1902–1904	32.2	1.13	0.015		
Miller	1921	32.0	1.12	0.08		
Miller	1923–1924	32.0	1.12	0.03		
Miller (Sunlight)	1924	32.0	1.12	0.014		
Tomascheck (Starlight)	1924	8.6	0.3	0.02		
Miller	1925–1926	32.0	1.12	0.088		
Mt Wilson)	1926	2.0	0.07	0.002		
Illingworth	1927	2.0	0.07	0.0002	0.0006	1 km/s
Piccard and Stahel (Rigi)	1927	2.8	0.13	0.006		
Michelson et al.	1929	25.9	0.9	0.01		
Joos	1930	21.0	0.75	0.002		

In recent times versions of the MM experiment have become commonplace. Lasers and masers amplify light by repeatedly bouncing it back and forth inside a carefully tuned cavity, thereby inducing high-energy atoms in the cavity to give off more light. The result is an effective path length of kilometers. Better yet, the light emitted in one cavity can be used to start the same cascade in another set at right angles, thereby creating an interferometer of extreme accuracy.

The first such experiment was led by Charles H. Townes, one of the co-creators of the first maser. Their 1958 experiment put an upper limit on drift, including any possible experimental errors, of only 30 m/s. In 1974 a repeat with accurate lasers in the triangular Trimmer experiment reduced this to 0.025 m/s, and included tests of entrainment by placing one leg in glass. In 1979 the Brillet-Hall experiment put an upper limit of 30 m/s for any one direction, but reduced this to only 0.000001 m/s for a two-direction case (ie, still or partially entrained aether). A year long repeat known as Hils and Hall, published in 1990, reduced this to $2x10^{-13}$.

Fallout

This result was rather astounding and not explainable by the then-current theory of wave propagation in a static aether. Several explanations were attempted, among them, that the experiment had a hidden flaw (apparently Michelson's initial belief), or that the Earth's gravitational field somehow "dragged" the aether around with it in such a way as locally to eliminate its effect. Miller would have argued that, in most if not all experiments other than his own, there was little possibility of detecting an aether wind since it was almost completely blocked out by the laboratory walls or by the apparatus itself. Be this as it may, the idea of a simple aether, what became known as the *First Postulate*, had been dealt a serious blow.

A number of experiments were carried out to investigate the concept of aether dragging, or *entrainment*. The most convincing was carried out by Hamar, who placed one arm of the interferometer between two huge lead blocks. If aether were dragged by mass, the blocks would, it was theorised, have been enough to cause a visible effect. Once again, no effect was seen.

Walter Ritz's Emission theory (or ballistic theory), was also consistent with the results of the experiment, not requiring aether, more intuitive and paradox-free. This became known as the *Second Postulate*. However it also led to several "obvious" optical effects that were not seen in astronomical photographs, notably in observations of binary stars in which the light from the two stars could be measured in an interferometer.

The Sagnac experiment placed the MM apparatus on a constantly rotating turntable. In doing so any ballistic theories such as Ritz's could be tested directly, as the light going one way around the device would have different length to travel than light going the other way (the eyepiece and mirrors would be moving toward/away from the light). In Ritz's theory there would be no shift, because the net velocity between the light source and detector was zero (they were both mounted on the turntable). However in this case an effect *was* seen, thereby eliminating any simple ballistic theory. This fringe-shift effect is used today in laser gyroscopes.

Another possible solution was found in the Lorentz-FitzGerald contraction hypothesis. In this theory all objects physically contract along the line of motion relative to the aether, so while the light may indeed transit slower on that arm, it also ends up travelling a shorter distance that exactly cancels out the drift.

In 1932 the Kennedy-Thorndike experiment modified the Michelson-Morley experiment by making the path lengths of the split beam unequal, with one arm being very long. In this version the two ends of the experiment were at different velocities due to the rotation of the earth, so the contraction would not "work out" to exactly cancel the result. Once again, no effect was seen.

Ernst Mach was among the first physicists to suggest that the experiment actually amounted to a disproof of the aether theory. The development of what became Einstein's special theory of relativity had the Fitzgerald-Lorentz contraction derived from the invariance postulate, and was also consistent with the apparently null results of most experiments (though not, as was recognised at the 1928 meeting, with Miller's observed seasonal effects). Today relativity is generally considered the "solution" to the MM null result.

The Trouton-Noble experiment is regarded as the electrostatic equivalent of the Michelson-Morley optical experiment, though whether or not it can ever be done with the necessary sensitivity is debatable. On the other hand, the 1908 Trouton-rankine experiment that spelled the end of the Lorentz-FitzGerald contraction hypothesis achieved an incredible sensitivity.

References

- W. Ritz, Recherches Critiques sur l'Electrodynamique Generale, *Ann. Chim.*, *Phys*., 13, 145, (1908) - English Translation
- W. de Sitter, Ein astronomischer Bewis für die Konstanz der Lichgeshwindigkeit, *Physik. Zeitschr*, 14, 429 (1913)

Mathematical approach

Introduction

The teaching of Special Relativity on undergraduate physics courses involves a considerable mathematical background knowledge. Particularly important are the manipulation of vectors and matrices and an elementary knowledge of curvature. The background mathematics is given below and can be skipped by those who are familiar with these techniques.

Vectors

Physical effects involve things acting on other things to produce a change of position, tension etc. These effects usually depend upon the strength, angle of contact, separation etc of the interacting things rather than on any absolute reference frame so it is useful to describe the rules that govern the interactions in terms of the relative positions and lengths of the interacting things rather than in terms of any fixed viewpoint or coordinate system. Vectors were introduced in physics to allow such relative descriptions.

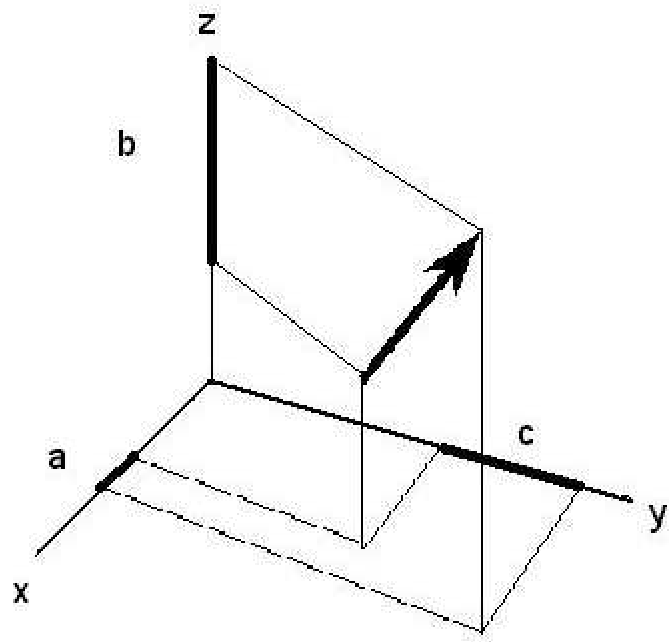
The use of vectors in elementary physics often avoids any real understanding of what they are. They are a new concept, as unique as numbers themselves, which have been related to the rest of mathematics and geometry by a series of formulae such as linear combinations, scalar products etc.

Vectors are defined as "directed line segments" which means they are lines drawn in a particular direction. The introduction of time as a geometric entity means that this definition of a vector is rather archaic, a better definition might be that a vector is information arranged as a continuous succession of points in space and time. Vectors have length and direction, the direction being from earlier to later.

Vectors are represented by lines terminated with arrow symbols to show the direction. A point that moves from the left to the right for about three centimetres can be represented as:

Symbol for a vector

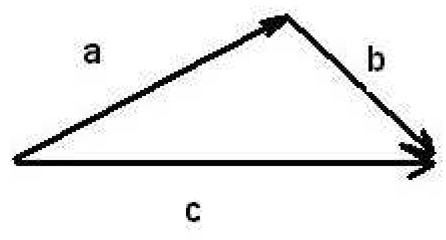
If a vector is represented within a coordinate system it has components along each of the axes of the system. These components do not normally start at the origin of the coordinate system.



The vector represented by the bold arrow has components a, b and c which are lengths on the coordinate axes. If the vector starts at the origin the components become simply the coordinates of the end point of the vector and the vector is known as the position vector of the end point.

Addition of Vectors

If two vectors are connected so that the end point of one is the start of the next the sum of the two vectors is defined as a third vector drawn from the start of the first to the end of the second:



c is the sum of **a** and **b**:

c = **a** + **b**

If a components of **a** are a, b, c and the components of **b** are d, e, f then the components of the sum of the two vectors are (a+d), (b+e) and (c+f). In other words, when vectors are added it is the components that add numerically rather than the lengths of the vectors themselves.

Rules of Vector Addition

1. Commutativity **a** + **b** = **b** + **a**

2. Associativity (**a** + **b**) + **c** = **a** + (**b** + **c**)

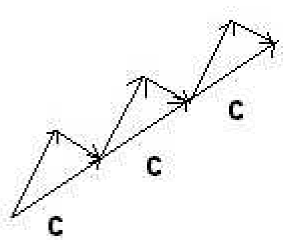
If the zero vector (which has no length) is labelled as 0

3. **a** + (-**a**) = 0

4. **a** + 0 = **a**

Rules of Vector Multiplication by a Scalar

The discussion of components and vector addition shows that if vector **a** has components a,b,c then q**a** has components qa, qb, qc. The meaning of vector multiplication is shown below:



The bottom vector c is added three times which is equivalent to multiplying it by 3.

1. Distributive laws q(**a** + **b**) = q**a** + q**b** and (q + p)**a** = q**a** + p**a**

2. Associativity q(p**a**) = qp**a**

Also 1 **a** = **a**

If the rules of vector addition and multiplication by a scalar apply to a set of elements they are said to define a vector space.

Linear Combinations and Linear Dependence

An element of the form:

$$q_1\mathbf{a_1} + q_2\mathbf{a_2} + q_3\mathbf{a_3} + ... + q_m\mathbf{a_m}$$

is called a linear combination of the vectors.

The set of vectors multiplied by scalars in a linear combination is called the span of the vectors. The word span is used because the scalars (q) can have any value - which means that any point in the subset of the vector space defined by the span can contain a vector derived from it.

Suppose there were a set of vectors (<math \mathbf{a_1,a_2,.... ,a_m</math>) , if it is possible to express one of these vectors in terms of the others, using any linear combination, then the set is said to be linearly dependent. If it is not possible to express any one of the vectors in terms of the others, using any linear combination, it is said to be linearly independent.

In other words, if there are values of the scalars such that:

(1). $\mathbf{a_1} = q_2\mathbf{a_2} + q_3\mathbf{a_3} + ... + q_m\mathbf{a_m}$

the set is said to be linearly dependent.

There is a way of determining linear dependence. From (1) it can be seen that if q_1 is set to minus one then:

$$q_1\mathbf{a_1} + q_2\mathbf{a_2} + q_3\mathbf{a_3} + ... + q_m\mathbf{a_m} = 0$$

So in general, if a linear combination can be written that sums to a zero vector then the set of vectors ($\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_m}$) are not linearly independent.

If two vectors are linearly dependent then they lie along the same line (wherever a and b lie on the line, scalars can be found to produce a linear combination which is a zero vector). If three vectors are linearly dependent they lie on the same line or on a plane (collinear or coplanar).

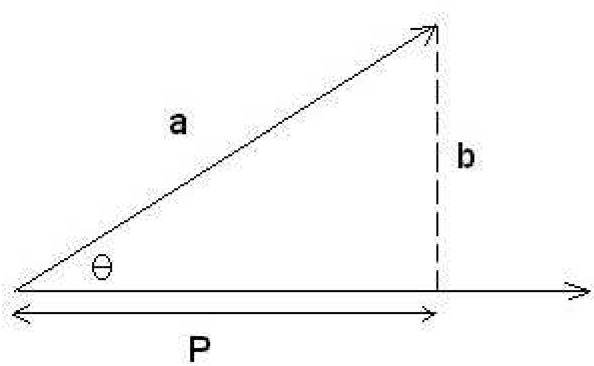
Dimension

If n+1 vectors in a vector space are linearly dependent then n vectors are linearly independent and the space is said to have a dimension of n. The set of n vectors is said to be the basis of the vector space.

Scalar Product

Also known as the 'dot product' or 'inner product'. The scalar product is a way of removing the problem of angular measures from the relationship between vectors and, as Weyl put it, a way of comparing the lengths of vectors that are arbitrarily inclined to each other.

Consider two vectors with a common origin:



The projection of $\mathbf{a}$ on $\mathbf{b}$ is:

$$P = |\mathbf{a}| cosq$$

Where $|\mathbf{a}|$ is the length of $\mathbf{a}$.

The scalar product is defined as:

(2) $\mathbf{a}.\mathbf{b} = |\mathbf{a}||\mathbf{b}|cosq$

Notice that cos q is zero if $\mathbf{a}$and $\mathbf{b}$are perpendicular. This means that if the scalar product is zero the vectors composing it are orthogonal (perpendicular to each other).

(2) also allows cos q to be defined as:

$$cosq = \mathbf{a}.\mathbf{b}/(|\mathbf{a}||\mathbf{b}|)$$

The definition of the scalar product also allows a definition of the length of a vector in terms of the concept of a vector itself. The scalar product of a vector with itself is:

$$\mathbf{a}.\mathbf{a} = |\mathbf{a}||\mathbf{a}|cos0$$

cos 0 (the cosine of zero) is one so:

$$\mathbf{a}.\mathbf{a} = a^2$$

which is our first direct relationship between vectors and scalars. This can be expressed as:

(3) $a = (\mathbf{a}.\mathbf{a})1/2$

where a is the length of $\mathbf{a}$.

Properties:

1. Linearity $[G\mathbf{a} + H\mathbf{b}].\mathbf{c} = G\mathbf{a}.\mathbf{c} + H\mathbf{b}.\mathbf{c}$

2. symmetry $\mathbf{a}.\mathbf{b} = \mathbf{b}.\mathbf{a}$

3. Positive definiteness $\mathbf{a}.\mathbf{a}$is greater than or equal to 0

4. Distributivity for vector addition $(\mathbf{a} + \mathbf{b}).\mathbf{c} = \mathbf{a}.\mathbf{c} + \mathbf{b}.\mathbf{c}$

5. Schwarz inequality $|\mathbf{a}.\mathbf{b}| \leq ab$

6. Parallelogram equality $|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = 2(|\mathbf{a}|^2 + |\mathbf{b}|^2)$

From the point of view of vector physics the most important property of the scalar product is the expression of the scalar product in terms of coordinates.

7. $\mathbf{a}.\mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3$

This gives us the length of a vector in terms of coordinates (Pythagoras' theorem) from:

8. $\mathbf{a}.\mathbf{a} = a_2 = a_1^2 + a_2^2 + a_3^2$

The derivation of 7 is:

$$\mathbf{a} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$$

where $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are unit vectors along the coordinate axes. From (4)

$$\mathbf{a}.\mathbf{b} = (a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}).\mathbf{b} = a_1\mathbf{i}.\mathbf{b} + a_2\mathbf{j}.\mathbf{b} + a_3\mathbf{k}.\mathbf{b}$$

but $\mathbf{b} = b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}$

so:

$$\mathbf{a}.\mathbf{b} = b_1a_1\mathbf{i}.\mathbf{i} + b_2a_1\mathbf{i}.\mathbf{j} + b_3a_1\mathbf{i}.\mathbf{k} + b_1a_2\mathbf{j}.\mathbf{i} + b_2a_2\mathbf{j}.\mathbf{j} + b_3a_2\mathbf{j}.\mathbf{k} + b_1a_3\mathbf{k}.\mathbf{i} + b_2a_3\mathbf{k}.\mathbf{j} + b_3a_3\mathbf{k}.\mathbf{k}$$

$\mathbf{i}.\mathbf{j}, \mathbf{i}.\mathbf{k}, \mathbf{j}.\mathbf{k},$ etc. are all zero because the vectors are orthogonal, also $\mathbf{i}.\mathbf{i}, \mathbf{j}.\mathbf{j}$ and $\mathbf{k}.\mathbf{k}$ are all one (these are unit vectors defined to be 1 unit in length).

Using these results:

$$\mathbf{a}.\mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3$$

Matrices

Matrices are sets of numbers arranged in a rectangular array. They are especially important in linear algebra because they can be used to represent the elements of linear equations.

11a + 2b = c

5a + 7b = d

The constants in the equation above can be represented as a matrix:

$$\mathbf{A} = \begin{bmatrix} 11 & 2 \\ 5 & 7 \end{bmatrix}$$

The elements of matrices are usually denoted symbolically using lower case letters:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Matrices are said to be equal if all of the corresponding elements are equal.

Eg: if $a_{ij} = b_{ij}$

Then $\mathbf{A} = \mathbf{B}$

Matrix Addition

Matrices are added by adding the individual elements of one matrix to the corresponding elements of the other matrix.

$c_{ij} = a_{ij} + b_{ij}$

or $\mathbf{C} = \mathbf{A} + \mathbf{B}$

Matrix addition has the following properties:

1. Commutativity $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$

2. Associativity $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$

and

3. $\mathbf{A} + (-\mathbf{A}) = 0$

4. $\mathbf{A} + 0 = \mathbf{A}$

From matrix addition it can be seen that the product of a matrix $\mathbf{A}$ and a number p is simply $p\mathbf{A}$ where every element of the matrix is multiplied individually by p.

Transpose of a Matrix

A matrix is transposed when the rows and columns are interchanged:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$\mathbf{A^T} = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

Notice that the principal diagonal elements stay the same after transposition.

A matrix is symmetric if it is equal to its transpose eg: $a_{kj} = a_{jk}$.

It is skew symmetric if $\mathbf{A^T} = -\mathbf{A}$ eg: $a_{kj} = -a_{jk}$. The principal diagonal of a skew symmetric matrix is composed of elements that are zero.

Other Types of Matrix

Diagonal matrix: all elements above and below the principal diagonal are zero.

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Unit matrix: denoted by I, is a diagonal matrix where all elements of the principal diagonal are 1.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Matrix Multiplication

Matrix multiplication is defined in terms of the problem of determining the coefficients in linear transformations.

Consider a set of linear transformations between 2 coordinate systems that share a common origin and are related to each other by a rotation of the coordinate axes.

Two Coordinate Systems Rotated Relative to Each Other

If there are 3 coordinate systems, x, y, and z these can be transformed from one to another:

$x_1 = a_{11}y_1 + a_{12}y_2$

$x_2 = a_{21}y_1 + a_{22}y_2$

$y_1 = b_{11}z_1 + b_{12}z_2$

$y_2 = b_{21}z_1 + b_{22}z_2$

$x_1 = c_{11}z_1 + c_{12}z_2$

$x_2 = c_{21}z_1 + c_{22}z_2$

By substitution:

$x_1 = a_{11}(b_{11}z_1 + b_{12}z_2) + a_{12}(b_{21}z_1 + b_{22}z_2)$

$x_2 = a_{21}(b_{11}z_1 + b_{12}z_2) + a_{22}(b_{21}z_1 + b_{22}z_2)$

$x_1 = (a_{11}b_{11} + a_{12}(b_{21})z_1 + (a_{11}b_{12} + a_{12}b_{22})z_2$

$x_2 = (a_{21}b_{11} + a_{22}(b_{21})z_1 + (a_{21}b_{12} + a_{22}b_{22})z_2$

Therefore:

$c_{11} = (a_{11}b_{11} + a_{12}(b_{21})$

$c_{12} = (a_{11}b_{12} + a_{12}b_{22})$

$c_{21} = (a_{21}b_{11} + a_{22}b_{21})$

$c_{22} = (a_{21}b_{12} + a_{22}b_{22})$

The coefficient matrices are:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

From the linear transformation the product of A and B is defined as:

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} (a_{11}b_{11} + a_{12}b_{21}) & (a_{11}b_{12} + a_{12}b_{22}) \\ (a_{21}b_{11} + a_{22}b_{21}) & (a_{21}b_{12} + a_{22}b_{22}) \end{bmatrix}$$

In the discussion of scalar products it was shown that, for a plane the scalar product is calculated as: $\mathbf{a}.\mathbf{b} = a_1b_1 + a_2b_2$ where a and b are the coordinates of the vectors a and b.

Now mathematicians define the rows and columns of a matrix as vectors:

A Column vector is $\mathbf{b} = \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix}$

And a Row vector $\mathbf{a} = \begin{bmatrix} a_{11} & a_{12} \end{bmatrix}$

Matrices can be described as vectors eg:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{a_1} \\ \mathbf{a_2} \end{bmatrix}$$

and

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{b_1} \mathbf{b_2} \end{bmatrix}$$

Matrix multiplication is then defined as the scalar products of the vectors so that:

$$\mathbf{C} = \begin{bmatrix} \mathbf{a_1}.\mathbf{b_1} & \mathbf{a_1}.\mathbf{b_2} \\ \mathbf{a_2}.\mathbf{b_1} & \mathbf{a_2}.\mathbf{b_2} \end{bmatrix}$$

From the definition of the scalar product $\mathbf{a_1}.\mathbf{b_1} = a_{11}b_{11} + a_{12}b_{21}$ etc.

In the general case:

$$\mathbf{C} = \begin{bmatrix} \mathbf{a_1.b_1} & \mathbf{a_1.b_2} & . & \mathbf{a_1.b_n} \\ \mathbf{a_2.b_1} & \mathbf{a_2.b_2} & . & \mathbf{a_2.b_n} \\ . & . & . & . \\ \mathbf{a_m.b_1} & \mathbf{a_m.b_2} & . & \mathbf{a_m.b_n} \end{bmatrix}$$

This is described as the multiplication of rows into columns (eg: row vectors into column vectors). The first matrix must have the same number of columns as there are rows in the second matrix or the multiplication is undefined.

After matrix multiplication the product matrix has the same number of rows as the first matrix and columns as the second matrix:

$\begin{bmatrix} 1 & 3 & 4 \\ 6 & 3 & 2 \end{bmatrix}$ times $\begin{bmatrix} 2 \\ 3 \\ 7 \end{bmatrix}$ has 2 rows and 1 column $\begin{bmatrix} 39 \\ 35 \end{bmatrix}$

ie: first row is 1 * 2 + 3 * 3 + 4 * 7 = 39 and second row is 6 * 2 + 3 * 3 + 2 * 7 = 35

$\mathbf{AB} = \begin{bmatrix} 1 & 3 & 2 \\ 2 & -1 & 3 \end{bmatrix}$ times $\begin{bmatrix} 2 & 3 & 4 \\ 3 & 2 & 1 \\ 5 & 1 & 3 \end{bmatrix}$ has 2 rows and 3 columns $\begin{bmatrix} 21 & 11 & 13 \\ 16 & 7 & 16 \end{bmatrix}$

Notice that $\mathbf{BA}$ cannot be determined because the **number of columns in the first matrix must equal the number of rows in the second matrix** to perform matrix multiplication.

Properties of Matrix Multiplication

1. Not commutative $\mathbf{AB} \neq \mathbf{BA}$

2. Associative $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$

$(k\mathbf{A})\mathbf{B} = k(\mathbf{AB}) = \mathbf{A}(k\mathbf{B})$

3. Distributative for matrix addition

$(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$

matrix multiplication is not commutative so $\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$ is a separate case.

4. The cancellation law is not always true:

$\mathbf{AB} = 0$ does not mean $\mathbf{A} = 0$ or $\mathbf{B} = 0$

There is a case where matrix multiplication is commutative. This involves the scalar matrix where the values of the principle diagonal are all equal. Eg:

$$\mathbf{S} = \begin{bmatrix} k & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & k \end{bmatrix}$$

In this case $\mathbf{AS} = \mathbf{SA} = k\mathbf{A}$. If the scalar matrix is the unit matrix: $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$.

Linear Transformations

A simple linear transformation such as:

$x_1 = a_{11}y_1 + a_{12}y_2$

$x_2 = a_{21}y_1 + a_{22}y_2$

can be expressed as:

$$\mathbf{x} = \mathbf{Ay}$$

eg:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} * \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

and

$y_1 = b_{11}z_1 + b_{12}z_2$

$y_2 = b_{21}z_1 + b_{22}z_2$

as: $\mathbf{y} = \mathbf{Bz}$

Using the associative law:

$$\mathbf{x} = \mathbf{A}(\mathbf{Bz}) = \mathbf{ABz} = \mathbf{Cz}$$

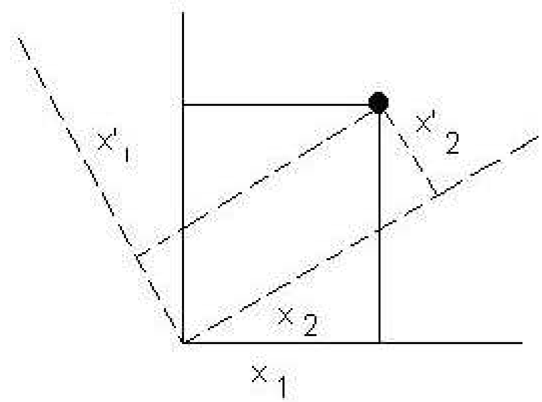
and so:

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} (a_{11}b_{11} + a_{12}b_{21}) & (a_{11}b_{12} + a_{12}b_{22}) \\ (a_{21}b_{11} + a_{22}b_{21}) & (a_{21}b_{12} + a_{22}b_{22}) \end{bmatrix}$$

as before.

Indicial Notation

Consider a simple rotation of coordinates:



x^μ is defined as x_1 , x_2

x^ν is defined as $x_1^{'}$, $x_2^{'}$

The scalar product can be written as:

$$\mathbf{s}.\mathbf{s} = g_{\mu\nu}x^\mu x^\nu$$

Where:

$$g_{\mu\nu} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and is called the metric tensor for this 2D space.

$$\mathbf{s}.\mathbf{s} = g_{11}x_1x_1^{'} + g_{12}x_1x_2^{'} + g_{21}x_2x_1^{'} + g_{22}x_2x_2^{'}$$

Now, $g_{11} = 1$, $g_{12} = 0$, $g_{21} = 0$, $g_{22} = 1$ so:

$$\mathbf{s}.\mathbf{s} = x_1x_1^{'} + x_2x_2^{'}$$

If there is no rotation of coordinates the scalar product is:

$$\mathbf{s}.\mathbf{s} = x_1x_1 + x_2x_2$$

$$s^2 = x_1^2 + x_2^2$$

Which is Pythagoras' theorem.

The Summation Convention

Indexes that appear as both subscripts and superscripts are summed over.

$$g_{\mu\nu}x^{\mu}x^{\nu} = g_{11}x_1x_1^{'} + g_{12}x_1x_2^{'} + g_{21}x_2x_1^{'} + g_{22}x_2x_2^{'}$$

by promoting n to a superscript it is taken out of the summation ie:.

$$g_{\mu}^{\nu}x^{\mu}x^{\nu} = g_{11}x_1x_1^{'} + +g_{21}x_2x_1^{'}$$

where ν = 1

Matrix Multiplication in Indicial Notation

Consider:

Columns times rows:

$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ times $\begin{bmatrix} y_1 & y_2 \end{bmatrix}$ = $\begin{bmatrix} x_1y_1 & x_1y_2 \\ x_2y_1 & x_2y_2 \end{bmatrix}$

Matrix product $\mathbf{XY} = x_iy_j$ Where i = 1, 2 j = 1, 2

There being no summation the indexes are both subscripts.

Rows times columns: $\begin{bmatrix} x_1 & x_2 \end{bmatrix}$ times $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ = $\begin{bmatrix} x_1y_1 & x_2y_2 \end{bmatrix}$

Matrix product $\mathbf{XY} = \delta_{ij}x^iy^j$

Where δ_{ij} is known as Kronecker delta and has the value 0 when $i \neq j$ and 1 when $i = j$. It is the indicial equivalent of the unit matrix:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

There being summation one value of i is a subscript and the other a superscript.

A matrix in general can be specified by any of:

M_i^j , M_{ij} , M_j^i , M^{ij} depending on which subscript or superscript is being summed over.

Vectors in Indicial Notation

A vector can be expressed as a sum of basis vectors.

$$\mathbf{x} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + a_3\mathbf{e}_3$$

In indicial notation this is: $x = a^ie_i$

Linear Transformations in indicial notation

Consider $\mathbf{x} = \mathbf{A}\mathbf{y}$ where $\mathbf{A}$ is a coefficient matrix and $\mathbf{x}$ and $\mathbf{y}$ are coordinate matrices.

In indicial notation this is:

$$x^\mu = A^\mu_\nu x^\nu$$

which becomes:

$$x_1 = a_{11}x_1^{'} + a_{12}x_2^{'} + a_{13}x_3^{'}$$

$$x_2 = a_{21}x_1^{'} + a_{22}x_2^{'} + a_{23}x_3^{'}$$

$$x_3 = a_{31}x_1^{'} + a_{32}x_2^{'} + a_{33}x_3^{'}$$

The Scalar Product in indicial notation

In indicial notation the scalar product is:

$$\mathbf{x}.\mathbf{y} = \delta_{ij}x^i y^j$$

Analysis of curved surfaces and transformations

It became apparent at the start of the nineteenth century that issues such as Euclid's parallel postulate required the development of a new type of geometry that could deal with curved surfaces and real and imaginary planes. At the foundation of this approach is Gauss's analysis of curved surfaces which allows us to work with a variety of coordinate systems and displacements on any type of surface.

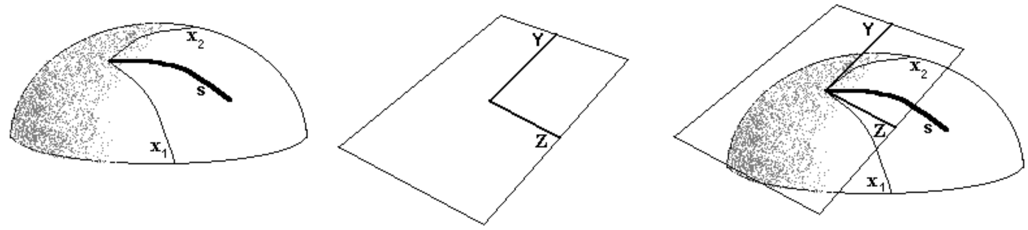
Elementary geometric analysis is useful as an introduction to Special Relativity because it suggests the physical meaning of the coefficients that appear in coordinate transformations.

Suppose there is a line on a surface. The length of this line can be expressed in terms of a coordinate system. A short length of line Δs in a two dimensional space may be expressed in terms of Pythagoras' theorem as:

$\Delta s^2 = \Delta x^2 + \Delta y^2$

Suppose there is another coordinate system on the surface with two axes: x_1, x_2, how can the length of the line be expressed in terms of these coordinates? Gauss tackled this problem and his analysis is quite straightforward for two coordinate axes:

Figure 1:



A curved surface with a set of coordinates on the surface

A plane with a set of Cartesian coordinates

The plane set at a tangent to the surface so that the coordinates can be compared

It is possible to use elementary differential geometry to describe displacements along the plane in terms of displacements on the curved surfaces:

$$\Delta Y = \Delta x_1 \frac{\delta Y}{\delta x_1} + \Delta x_2 \frac{\delta Y}{\delta x_2}$$

$$\Delta Z = \Delta x_1 \frac{\delta Z}{\delta x_1} + \Delta x_2 \frac{\delta Z}{\delta x_2}$$

The displacement of a short line is then assumed to be given by a formula, called a metric, such as Pythagoras' theorem

$\Delta S^2 = \Delta Y^2 + \Delta Z^2$

The values of ΔY and ΔZ can then be substituted into this metric:

$$\Delta S^2 = (\Delta x_1 \frac{\delta Y}{\delta x_1} + \Delta x_2 \frac{\delta Y}{\delta x_2})^2 + (\Delta x_1 \frac{\delta Z}{\delta x_1} + \Delta x_2 \frac{\delta Z}{\delta x_2})^2$$

Which, when expanded, gives the following:

$\Delta S^2 =$

$$(\frac{\delta Y}{\delta x_1}\frac{\delta Y}{\delta x_1} + \frac{\delta Z}{\delta x_1}\frac{\delta Z}{\delta x_1})\Delta x_1 \Delta x_1$$

$$+(\frac{\delta Y}{\delta x_2}\frac{\delta Y}{\delta x_1} + \frac{\delta Z}{\delta x_2}\frac{\delta Z}{\delta x_1})\Delta x_2 \Delta x_1$$

$$+(\frac{\delta Y}{\delta x_1}\frac{\delta Y}{\delta x_2} + \frac{\delta Z}{\delta x_1}\frac{\delta Z}{\delta x_2})\Delta x_1 \Delta x_2$$

$$+(\frac{\delta Y}{\delta x_2}\frac{\delta Y}{\delta x_2} + \frac{\delta Z}{\delta x_2}\frac{\delta Z}{\delta x_2})\Delta x_2 \Delta x_2$$

This can be represented using summation notation:

$$\Delta S^2 = \sum_{i=1}^{2}\sum_{k=1}^{2}(\frac{\delta Y}{\delta x_i}\frac{\delta Y}{\delta x_k} + \frac{\delta Z}{\delta x_i}\frac{\delta Z}{\delta x_k})\Delta x_i \Delta x_k$$

Or, using **indicial** notation:

$$\Delta S^2 = g_{ik}\Delta x^i \Delta x^k$$

Where:

$$g_{ik} = (\frac{\delta Y}{\delta x^i}\frac{\delta Y}{\delta x^k} + \frac{\delta Z}{\delta x^i}\frac{\delta Z}{\delta x^k})$$

If the coordinates are not merged then Δs is dependent on both sets of coordinates. In matrix notation:

$$\Delta s^2 = \mathbf{g}\Delta\mathbf{x}\Delta\mathbf{x}$$

becomes:

$$\begin{bmatrix}\Delta x_1 & \Delta x_2\end{bmatrix} \text{ times } \begin{bmatrix}a & b \\ c & d\end{bmatrix} \text{ times } \begin{bmatrix}\Delta x_1 \\ \Delta x_2\end{bmatrix}$$

Where a, b, c, d stand for the values of g_{ik}.

Therefore:

$$\begin{bmatrix}\Delta x_1 a + \Delta x_2 c & \Delta x_1 b + \Delta x_2 d\end{bmatrix} \text{ times } \begin{bmatrix}\Delta x_1 \\ \Delta x_2\end{bmatrix}$$

Which is:

$$(\Delta x_1 a + \Delta x_2 c)\Delta x_1 + (\Delta x_1 b + \Delta x_2 d)\Delta x_2 = \Delta {x_1}^2 a + 2\Delta x_1 \Delta x_2 (c+b) + \Delta {x_2}^2 d$$

So:

$$\Delta s^2 = \Delta {x_1}^2 a + 2\Delta x_1 \Delta x_2 (c+b) + \Delta {x_2}^2 d$$

Δs^2 is a **bilinear form** that depends on both Δx_1 and Δx_2. It can be written in matrix notation as:

$$\Delta s^2 = \mathbf{\Delta x^T A \Delta x}$$

Where A is the matrix containing the values in g_{ik}. This is a special case of the bilinear form known as the quadratic form because the same matrix ($\mathbf{\Delta x}$) appears twice; in the generalised bilinear form $\mathbf{B} = \mathbf{x^T A y}$ (the matrices $\mathbf{x}$ and $\mathbf{y}$ are different).

If the surface is a Euclidean plane then the values of gik are:

$$\begin{bmatrix} \delta Y/\delta x_1 \delta Y/\delta x_1 + \delta Z/\delta x_1 \delta Z/\delta x_1 & \delta Y/\delta x_2 \delta Y/\delta x_1 + \delta Z/\delta x_2 \delta Z/\delta x_1 \\ \delta Y/\delta x_2 \delta Y/\delta x_1 + \delta Z/\delta x_2 \delta Z/\delta x_1 & \delta Y/\delta x_2 \delta Y/\delta x_2 + \delta Z/\delta x_2 \delta Z/\delta x_2 \end{bmatrix}$$

Which become:

$$g_{\mu\nu} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

So the matrix A is the unit matrix I and:

$$\Delta s^2 = \mathbf{\Delta x^T I \Delta x}$$

and:

$$\Delta s^2 = \Delta x_1{}^2 + \Delta x_2{}^2$$

Which recovers Pythagoras' theorem yet again.

If the surface is derived from some other metric such as $\Delta s^2 = -\Delta Y^2 + \Delta Z^2$ then the values of gik are:

$$\begin{bmatrix} -\delta Y/\delta x_1 \delta Y/\delta x_1 + \delta Z/\delta x_1 \delta Z/\delta x_1 & -\delta Y/\delta x_2 \delta Y/\delta x_1 + \delta Z/\delta x_2 \delta Z/\delta x_1 \\ -\delta Y/\delta x_2 \delta Y/\delta x_1 + \delta Z/\delta x_2 \delta Z/\delta x_1 & -\delta Y/\delta x_2 \delta Y/\delta x_2 + \delta Z/\delta x_2 \delta Z/\delta x_2 \end{bmatrix}$$

Which becomes:

$$g_{\mu\nu} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

Which allows the original metric to be recovered ie: $\Delta s^2 = -\Delta x_1{}^2 + \Delta x_2{}^2$.

It is interesting to compare the geometrical analysis with the transformation based on matrix algebra that was derived in the section on indicial notation above:

$$\mathbf{s.s} = g_{11} x_1 x_1' + g_{12} x_1 x_2' + g_{21} x_2 x_1' + g_{22} x_2 x_2'$$

Now,

$$g_{\mu\nu} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

ie: $g_{11} = 1$, $g_{12} = 0$, $g_{21} = 0$, $g_{22} = 1$ so:

$$\mathbf{s.s} = x_1 x_1' + x_2 x_2'$$

If there is no rotation of coordinates the scalar product is:

$$\mathbf{s.s} = x_1 x_1 + x_2 x_2$$

$$s^2 = x_1^2 + x_2^2$$

Which recovers Pythagoras' theorem. However, the reader may have noticed that Pythagoras' theorem had been assumed from the outset in the derivation of the scalar product (see above).

The geometrical analysis shows that if a metric is assumed and the conditions that allow differential geometry are present then it is possible to derive one set of coordinates from another. This analysis can also be performed using matrix algebra with the same assumptions.

The example above used a simple two dimensional Pythagorean metric, some other metric such as the metric of a 4D Minkowskian space:

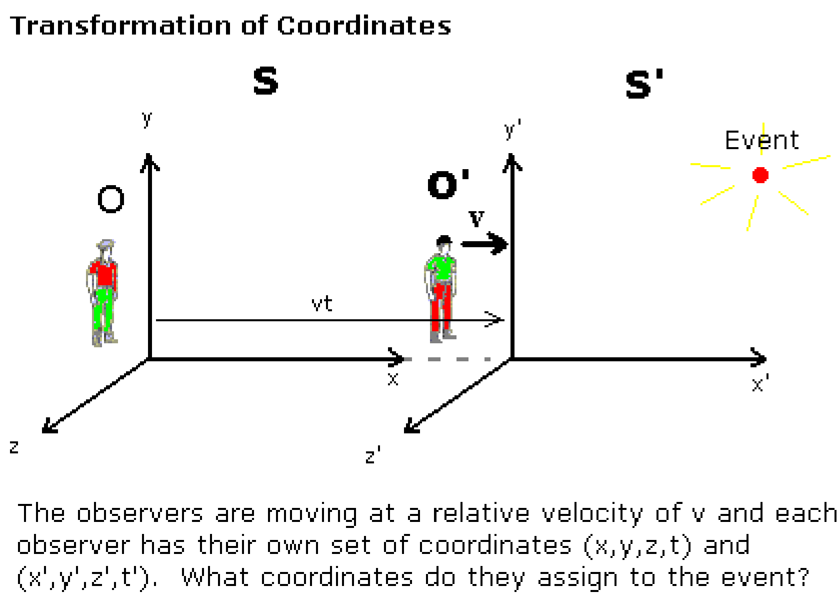
$$\Delta S^2 = -\Delta T^2 + \Delta X^2 + \Delta Y^2 + \Delta Z^2$$

could be used instead of Pythagoras' theorem.

Mathematical transformations

The Lorentz transformation

The Lorentz transformation deals with the problem of observers who are moving relative to each other. How are the coordinates of an event recorded by one observer related to the coordinates of the event recorded by the other observer? The **standard configuration** used in the calculation of the Lorentz transformation is shown below:



There are several ways of deriving the Lorentz transformations. The usual method is to work from Einstein's postulates (that the laws of physics are the same between all inertial reference frames and the speed of light is constant) whilst adding assumptions about isotropy, linearity and homogeneity. The second is to work from the assumption of a four dimensional Minkowskian metric.

In mathematics transformations are frequently symbolised with the "maps to" symbol:

$$(x, y, z, t) \mapsto (x', y', z', t')$$

The linearity and homogeneity of spacetime

Consider a clock moving freely, according to Newton's first law, that objects continue in a state of uniform motion unless acted upon by a force, the velocity of the clock in any given direction (dx_i / dt) is a constant.

If the clock is a real clock with readings given by τ then the relationship between these readings and the elapsed time anywhere in an inertial frame of reference, $dt / d\tau$, will be a constant. If the clock were to tick at an uneven rate compared with other clocks then the universe would not be homogenous in time - at some times the clock would appear to accelerate. This would also mean that Newton's first law would be broken and the universe would not be homogenous in space.

If dx_i / dt and $dt / d\tau$ are constant then $dx_\mu(\mu = 1,2,3,4)$ is also constant. This means that the clock is not accelerating ie: $d^2x_\mu / d\tau^2 = 0$.

Linearity is demonstrated by the way that the length of things does not depend on position or relative position; for instance, if $x' = ax^2$ the distance between two points would depend upon the position of the observer whereas if the relationship is linear ($x' = ax$) separations are independent of position.

The linearity and homogeneity assumptions mean that the coordinates of objects in the S' inertial frame are related to those in the S inertial frame by:

$$x'_\nu = (\sum \Lambda_{\nu\mu} x_\mu) + B_\nu$$

This formula is known as a **poincare transformation**. It can be expressed in indicial notation as:

$$x'^\nu = \Lambda^\nu_\mu x^\mu + B^\nu$$

If the origins of the frames coincide then B_ν can be assumed to be zero and the equation:

$$x'_\nu = \sum \Lambda_{\nu\mu} x_\mu$$

Those who are unfamiliar with the notation should note that the symbols x_1 etc. mean $x_1 = x, x_2 = y, x_3 = z, x_4 = t$ so the equation above is shorthand for:

$x' = a_{11}x + a_{12}y + a_{13}z + a_{14}t$

$y' = a_{21}x + a_{22}y + a_{23}z + a_{24}t$

$z' = a_{31}x + a_{32}y + a_{33}z + a_{34}t$

$t' = a_{41}x + a_{42}y + a_{43}z + a_{44}t$

In matrix notation the set of equations can be written as:

$$\mathbf{x}' = \mathbf{\Lambda x}$$

The **standard configuration** (see diagram above) has several properties, for instance:

The spatial origin of both observer's coordinate systems lies on the line of motion so the x axes can be chosen to be parallel.

The point given by $x = vt$ is the same as $x' = 0$.

The origins of both coordinate systems can coincide so that clocks can be synchronised when they are next to each other.

The coordinate planes ,y, y' and z,z', can be arranged to be orthogonal (at right angles) to the direction of motion.

Isotropy means that coordinate planes that are orthogonal at y=0 and z=0 in one frame are orthogonal at at y'=0 and z'=0 in the other frame.

According to the relativity principle any transformations between the same two inertial frames of reference must be the same. This is known as the **reciprocity theorem**.

The Lorentz transformation

From the linearity assumption and given that at the origin $y = 0 = y'$ so there is no constant offset then $y' = Ky$ and $y = Ky'$, therefore K=1. So:

$y' = y$

and, by the same reasoning:

$z' = z$

Now, considering the x coordinate of the event, the x and y axes can be assumed to be 0 (ie: an arbitrary shift of the coordinates to allow the event to lie on the x axes). If this is done then the linearity consideration and the fact that $x = vt$ and $x' = 0$ are the same point gives:

(1) $x' = \gamma(x - vt)$

where γ is a constant. According to the reciprocity theorem we also have:

(2) $x = \gamma(x' + vt')$

Einstein's assumption that the speed of light is a constant can now be introduced so that $x = ct$ and also $x' = ct'$. So:

$ct' = \gamma t(c - v)$

and

$ct = \gamma t'(c + v)$

So:

$c^2tt' = \gamma^2tt'(c^2 - v^2)$

and

$$\gamma = \frac{1}{\sqrt{1 - \mathbf{v}^2/\mathbf{c}^2}}$$

Therefore the Lorentz transformation equations are:

$t' = \gamma(t - vx / c^2)$

$x' = \gamma(x - vt)$

$y' = y$

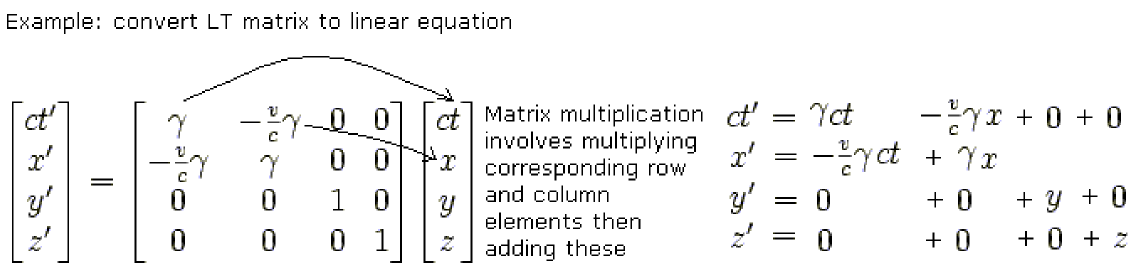
$z' = z$

The transformation for the time coordinate can derived from the transformation for the x coordinate assuming $x = ct$ and $x' = ct'$ or directly from equations (1) and (2) with a similar substitution for $x = ct$.

The coefficients of the Lorentz transformation can be represented in matrix format:

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \gamma & -\frac{v}{c}\gamma & 0 & 0 \\ -\frac{v}{c}\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix}.$$

A coordinate transformation of this type, that is due to motion, is known as a **boost**.



The Lorentz transformation equations can be used to show that:

$c^2dt'^2 - dx'^2 - dy'^2 - dz'^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$

Although whether the assumptions of linearity, isotropy and homogeneity in the derivation of the Lorentz transformation actually assumed this identity from the outset is a mute point.

Given that: $c^2dt'^2 - dx'^2 - dy'^2 - dz'^2$ also equals $c^2dt''^2 - dx''^2 - dy''^2 - dz''^2$ and a continuous range of other transformations it is clear that:

$$\Delta s^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$$

The quantity Δs is known as the **spacetime interval** and the quantity Δs^2 is known as the **squared displacement**.

A given squared displacement is constant for all observers, no matter how fast they are travelling, it is said to be **invariant**.

The equation:

$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$

is known as the **metric** of spacetime.

The geometry of space-time

The discussion above was simplified by assuming that the symbols x,y,z,t were to be understood as intervals. The treatment given below is suitable for an undergraduate level of presentation. SR uses a 'flat' 4-dimensional Minkowski space, which is an example of a space-time. This space, however, is very similar to the standard 3 dimensional Euclidean space, and fortunately by that fact, very easy to work with.

The differential of distance(ds) in cartesian 3D space is defined as:

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2$$

where (dx_1,dx_2,dx_3) are the differentials of the three spatial dimensions. In the geometry of special relativity, a fourth dimension, time, is added, with units of c, so that the equation for the differential of distance becomes:

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 - c^2dt^2$$

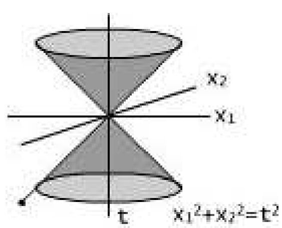
In many situations it may be convenient to treat time as imaginary (e.g. it may simplify equations), in which case t in the above equation is replaced by $i.t'$, and the metric becomes

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 + c^2(dt')^2$$

If we reduce the spatial dimensions to 2, so that we can represent the physics in a 3-D space

$$ds^2 = dx_1^2 + dx_2^2 - c^2dt^2$$

We see that things such as light which move at the speed of light lie along a dual-cone:



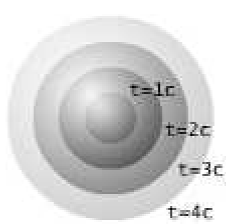
defined by the equation

$$ds^2 = 0 = dx_1^2 + dx_2^2 - c^2dt^2$$

or

$$dx_1^2 + dx_2^2 = c^2dt^2$$

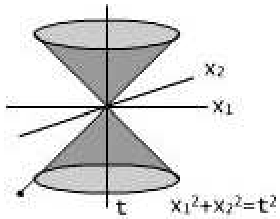
Which is the equation of a circle with $r=c*dt$. The path of something that moves at the speed of light is known as a **null geodesic**. If we extend the equation above to three spatial dimensions, the null geodesics are continuous concentric spheres, with radius = distance = c×(±time).



$$ds^2 = 0 = dx_1^2 + dx_2^2 + dx_3^2 - c^2dt^2$$

$$dx_1^2 + dx_2^2 + dx_3^2 = c^2 dt^2$$

This null dual-cone represents the "line of sight" of a point in space. That is, when we look at the stars and say "The light from that star which I am receiving is X years old.", we are looking down this line of sight: a null geodesic. We are looking at an event $d = \sqrt{x_1^2 + x_2^2 + x_3^2}$ meters away and *d/c* seconds in the past. For this reason the null dual cone is also known as the 'light cone'. (The point in the lower left of the picture below represents the star, the origin represents the observer, and the line represents the null geodesic "line of sight".)



The cone in the *-t* region is the information that the point is 'receiving', while the cone in the *+t* section is the information that the point is 'sending'.

Length contraction, time dilation and phase

Consider two inertial frames in standard configuration. There is a rigid rod moving along in the second frame at v m/s. The length of the rod is determined by observing the positions of the end points of the rod simultaneously - if the rod is moving it would be nonsense to use any other measure of length. An observer who is moving at the same velocity as the rod measures its "rest length". The Lorentz transformation for coordinates along the x axis is:

$x' = \gamma(x - vt)$

Suppose the positions, x_1,x_2, of the two ends of the rod are determined simultaneously (ie: *t* is constant):

$$(x_1^1 - x_2') = \gamma(x_1 - x_2)$$

Or, using $L_0 = (x_1' - x_2')$ for the rest length of the rod and $L = (x_1 - x_2)$ for the length of the rod that is measured by the observer who sees it fly past at v m/s:

$$\mathbf{L_0 = \gamma L}$$

Or, elaborating γ:

$$\mathbf{L = L_0\sqrt{1 - v^2/c^2}}$$

In other words the length of an object moving with velocity *v* is contracted in the direction of motion by a factor $\sqrt{1 - v^2/c^2}$ in the direction of motion.

The Lorentz transformation also affects the rate at which clocks appear to change their readings. The Lorentz transformation for time is:

$t' = \gamma(t - vx / c^2)$

This transformation has two components:

$t' = \gamma t - \gamma vx / c^2$

and is a straight line graph (ie: $t' = mt + c$).

The gradient of the graph is γ so:

$\Delta t' = \gamma\Delta t$

or:

$$t'_1 - t'_2 = \gamma(t_1 - t_2)$$

Therefore clocks in the moving frame will appear to go slow, if T_0 is a time interval in the rest frame and T is a time interval in the moving frame:

$T = \gamma T_0$

Or, expanding:

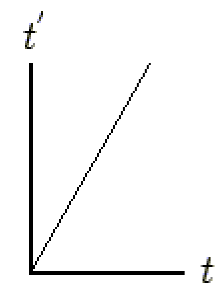
$$\mathbf{T} = \frac{\mathbf{T_0}}{\sqrt{\mathbf{1 - v^2/c^2}}}$$

The intercept of the graph is:

$\gamma vx / c^2$

This means that if a clock at point x is compared with a clock that was synchronised between frames at the origin it will show a constant time difference of $\gamma vx / c^2$ seconds. This quantity is known as the relativistic phase difference or "phase".

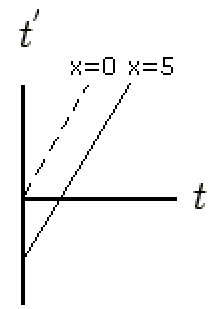
Time dilation and phase



The graph of:

$$t' = \gamma t - \gamma vx/c^2$$

For clocks synchronised at x=0 (the origin).

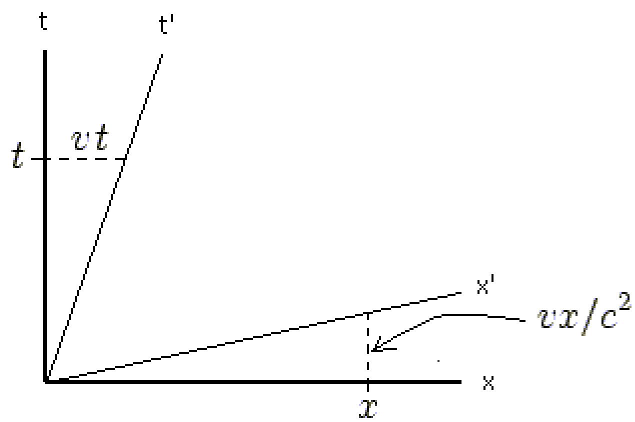


The graph of:

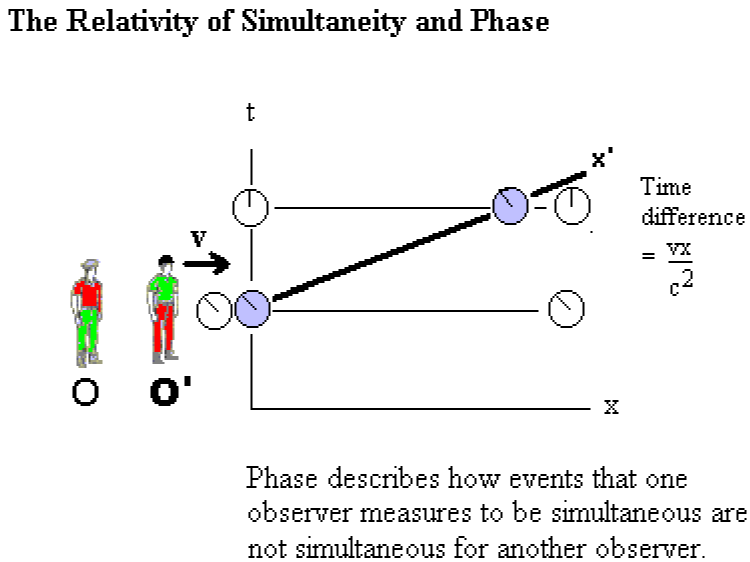
$$t' = \gamma t - \gamma vx/c^2$$

For clocks synchronised at x=5.

A comparison of the x and t axes in the standard configuration.



The relativistic phase is as important as the length contraction and time dilation results. It is the amount by which clocks that are synchronised at the origin go out of synchronisation with distance along the direction of travel. Phase affects all clocks except those at the point where clocks are syncronised and the infinitessimal y and z planes that cut this point. All clocks everywhere else will be out of synchronisation between the frames. The effect of phase is shown in the illustration below:



If the inertial frames are each composed of arrays of clocks spread over space then the clocks will be out of synchronisation as shown in the illustration above.

Hyperbolic geometry

In the flat spacetime of Special Relativity:

$s^2 = c^2t^2 - x^2 - y^2 - z^2$

Considering the x-axis alone:

$s^2 = c^2t^2 - x^2$

The standard equation of a hyperbola is:

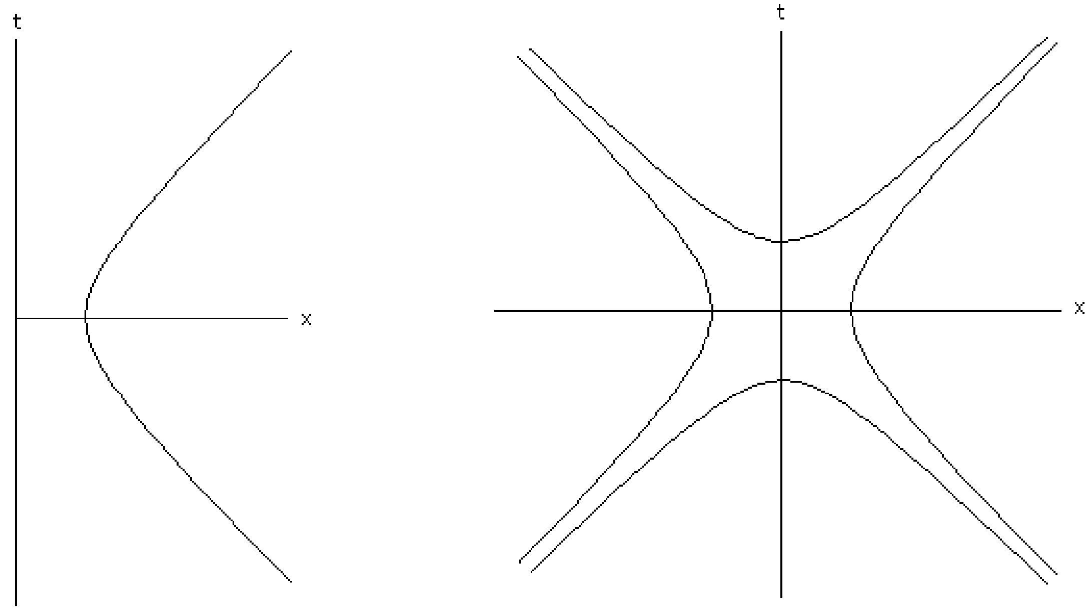
$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

In the case of spacetime:

$$\frac{(ct)^2}{s^2} - \frac{x^2}{s^2} = 1$$

Spacetime intervals separate one place or event in spacetime from another. So, for a given motion from one place to another or a given fixed length in one reference frame, given time interval etc. the metric of spacetime describes a hyperbolic space. This hyperbolic space encompasses the coordinates of all the observations made of the given interval by any observers.

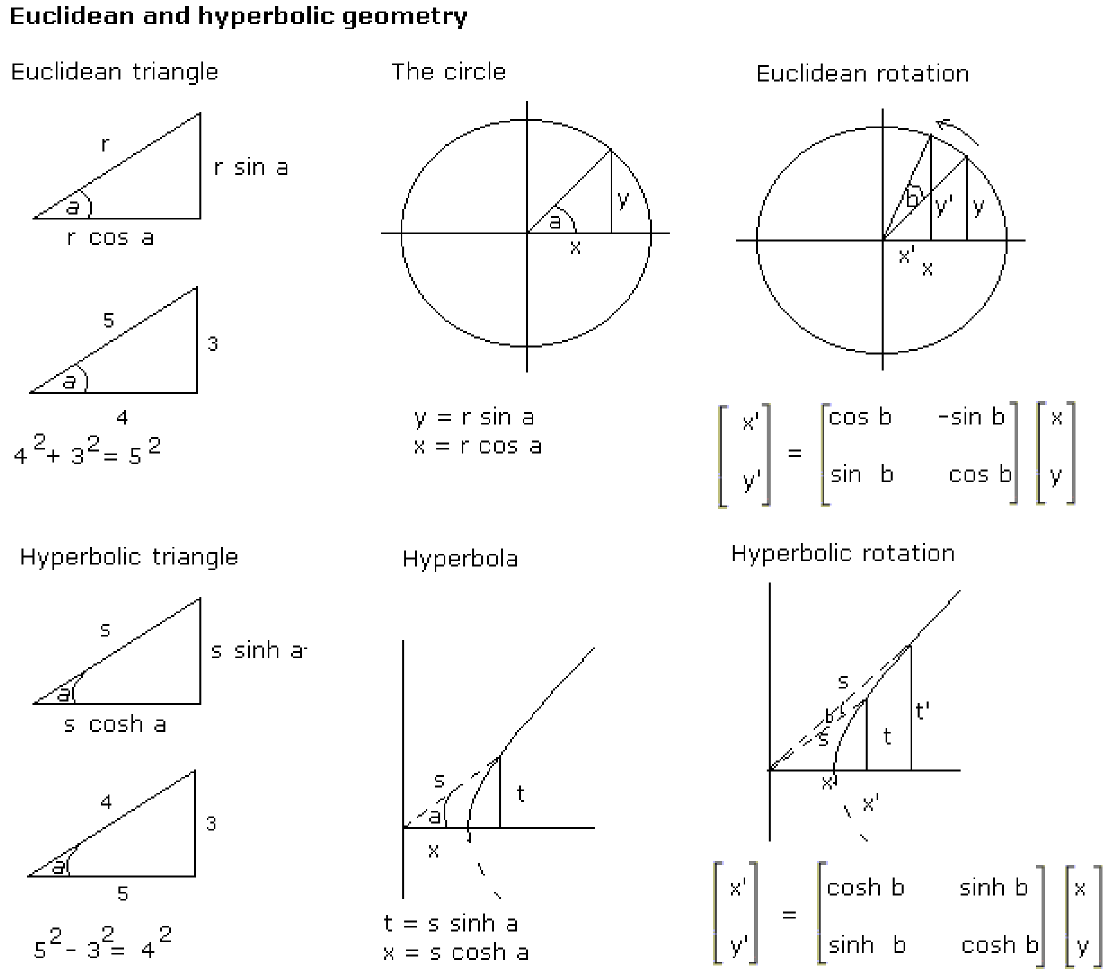
Hyperbolic spacetime



The general form of a hyperbola

A plot of the hyperbola that represents all values of t and x that observers might measure for a given spacetime interval.

It is possible to conceive of rotations in hyperbolic space in a similar way to rotations in Euclidean space. The idea of a rotation in hyperbolic space is summarised in the illustration below:



A rotation in hyperbolic space is equivalent to changing from one frame of reference to another whilst observing the same spacetime interval. It is moving from coordinates that give:

$(ct)^2 - x^2 = s^2$

to coordinates that give:

$$(ct')^2 - x'^2 = s^2$$

The formula for a rotation in hyperbolic space provides an alternative form of the Lorentz transformation ie:

$$\begin{bmatrix} ct \\ x \end{bmatrix} = \begin{bmatrix} \cosh\phi & \sinh\phi \\ \sinh\phi & \cosh\phi \end{bmatrix} \begin{bmatrix} ct' \\ x' \end{bmatrix}.$$

From which:

$x = x' \cosh\varphi + ct' \sinh\varphi$

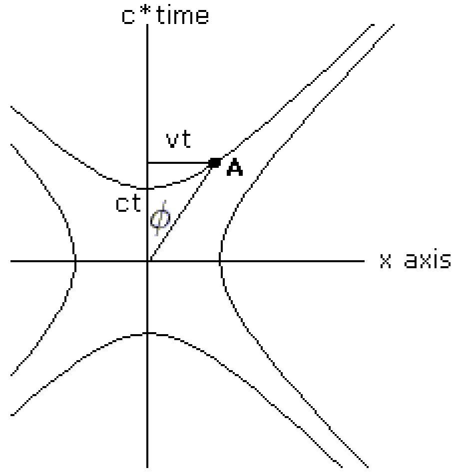
$ct = x' \sinh\varphi + ct' \cosh\varphi$

The value of φ can be determined by considering the coordinates assigned to a moving light that moves along the x axis from the origin at *vmsec*[1] flashes on for *t* seconds then flashes off.

The coordinates assigned by an observer on the light are: t',0,0,0, the coordinates assigned by the stationary observer are $t, x = vt$,0,0. The hyperbola representing these observations is illustrated below:

The value of ϕ from the equation

$$(ct)^2 - x^2 = s^2$$



tanh ϕ = vt/ct = v/c

The equation of the hyperbola is:

$(ct)^2 - x^2 = s^2 = (ct')^2$

but x=vt for the end of the flash so:

$$\tanh\phi = \frac{v}{c}$$

Now, from hyperbolic trigonometry:

$$\frac{1}{\sqrt{1-\tanh^2\phi}} = \sqrt{\frac{\cosh^2\phi}{\cosh^2\phi - \sinh^2\phi}} = \cosh\phi$$

But $\tanh\phi = \frac{v}{c}$ so:

$$\cosh\phi = \frac{1}{\sqrt{1 - v^2/c^2}} = \gamma$$

and, from the hyperbolic trigonometric formula sinhφ = tanhφcoshφ:

$$\sinh \phi = \frac{v}{c}\gamma$$

Inserting these values into the equations for the hyperbolic rotation:

$x = x^{'}\cosh\varphi + ct^{'}\sinh\varphi$

$x = \gamma x^{'} + ct^{'}\gamma v / c$

Which gives the standard transform for x:

$x = \gamma(x^{'} + vt^{'})$

In a similar way $ct = x^{'}\sinh\varphi + ct^{'}\cosh\varphi$ is equivalent to:

$t = \gamma(t^{'} + vx / c^2)$

So the Lorentz transformations can also be derived from the assumption that boosts are equivalent to rotations in hyperbolic space with a metric $s^2 = c^2t^2 - x^2 - y^2 - z^2$.

The quantity φ is known as the **rapidity** of the boost.

Addition of velocities

Suppose there are three observers 1, 2, and 3 who are moving at different velocities along the x-axis. Observers 1 and 2 are moving at a relative velocity v and observers 2 and 3 are moving at a relative velocity of $u^{'}$. The problem is to determine the velocity of observer 3 as observed by observer 1 (u).

It turns out that there is a very convenient relationship between rapidities that solves this problem:

If $v / c = \tanh\varphi$ and $u^{'} / c = \tanh\alpha$ then:

$u / c = \tanh(\varphi + \alpha)$

In other words the rapidities can be simply added from one observer to another ie:

$\sigma = \varphi + \alpha$

Hence:

$\tanh(\sigma) = \tanh(\varphi + \alpha)$

So the velocities can be added by simply adding the rapidities. Using hyperbolic trigonometry:

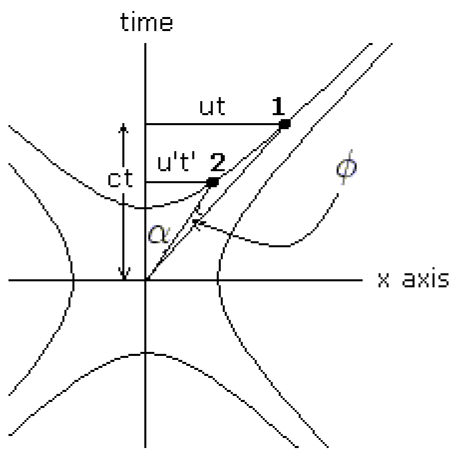
$$u/c = \tanh(\alpha + \phi) = \frac{\tanh \alpha + \tanh \phi}{1 + \tanh \alpha \tanh \phi} = \frac{u^{'}/c + v/c}{1 + u^{'}v/c^2}$$ Therefore:

$$\mathbf{u} = \frac{\mathbf{u}^{'} + \mathbf{v}}{\mathbf{1} + \mathbf{u}^{'}\mathbf{v}/\mathbf{c}^{\mathbf{2}}}$$

Which is the relativistic velocity addition theorem.

The relationship $u / c = \tanh(\varphi + \alpha)$ is shown below:

Relative velocities



$$u/c = \tanh(\alpha + \phi)$$

From the viewpoint of observer 3 observer 2 is separated by angle α and observer 1 is separated from observer 2 by angle ϕ .

Velocity transformations can be obtained without referring to the rapidity. The general case of the transformation of velocities in any direction is derived as follows:

$$\mathbf{u}' = (u_1', u_2', u_3')$$

where u_1' etc. are the components of the velocity in the x, y, z directions.

Writing out the components of velocity:

$$u_1' = dx'/dt'$$

$$u_2' = dy'/dt'$$

$$u_3' = dz'/dt'$$

But from the Lorentz transformations:

$dx' = \gamma(dx - vdt)$

$dy' = dy$

$dz' = dz$

$dt' = \gamma(dt - vdx / c^2)$

Therefore:

$$u_1' = dx'/dt' = \frac{\gamma(dx - v\,dt)}{\gamma(dt - v\,dx/c^2)}$$

$$u_2' = dy'/dt' = \frac{dy}{\gamma(dt - v\,dx/c^2)}$$

$$u_3' = dz'/dt' = \frac{dz}{\gamma(dt - v\,dx/c^2)}$$

Dividing top and bottom of each fraction by *dt*:

$$u_1' = \frac{\gamma(dx/dt - v)}{\gamma(1 - v\,dx/dt/c^2)}$$

$$u_2' = \frac{dy/dt}{\gamma(1 - v\,dx/dt/c^2)}$$

$$u_3' = \frac{dz/dt}{\gamma(1 - v\,dx/dt/c^2)}$$

Substituting $\mathbf{u} = (u_1, u_2, u_3)$

$$u_1' = \frac{u - v}{1 - uv/c^2}$$

$$u_2' = \frac{u_2}{\gamma(1 - uv/c^2)}$$

$$u_3' = \frac{u_3}{\gamma(1 - uv/c^2)}$$

The full velocity transformations are tabulated below:

$u'_x = \frac{(u_x - v)}{(1 - u_x v/c^2)}$	$u_x = \frac{(u'_x + v)}{(1 + u'_x v/c^2)}$
$u'_y = \frac{u_y\sqrt{1 - v^2/c^2}}{(1 - u_x v/c^2)}$	$u_y = \frac{u'_y\sqrt{1 - v^2/c^2}}{(1 + u'_x v/c^2)}$
$u'_z = \frac{u_z\sqrt{1 - v^2/c^2}}{(1 - u_x v/c^2)}$	$u_z = \frac{u'_z\sqrt{1 - v^2/c^2}}{(1 + u'_x v/c^2)}$

Having calculated the components of the velocity vector it it now possible to calculate the magnitudes of the overall vectors between frames:

$$u = \sqrt{u_1^2 + u_2^2 + u_3^2}$$

$$u' = \sqrt{u_1'^2 + u_2'^2 + u_3'^2}$$

Acceleration transformation

It was seen above that:

$u / c = \tanh\varphi$

and, if $v / c = \tanh\alpha$ and $u' / c = \tanh\varepsilon$ the velocity addition theorem can be expressed as the sum of the rapidities:

$\varphi = \alpha + \varepsilon$

If we differentiate this equation with respect to t to investigate acceleration, then assuming v is constant:

(1) $$\frac{d\phi}{dt} = \frac{d\epsilon}{dt'}\frac{dt'}{dt}$$

But $\frac{d\phi}{dt}$ is also equal to:

$$\frac{d\phi}{dt} = \frac{d\phi}{du}\frac{du}{dt}$$

But $\varphi = \tanh^{-1}(u / c)$ and the derivative of an arctangent is given by:

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

and hence:

$$\frac{d\phi}{du} = \frac{1}{c}\frac{1}{(1 - u^2/c^2)}$$

But:

$$\frac{1}{1 - u^2/c^2} = \gamma^2(u)$$

ie: $\gamma(u)$ is gamma for observers moving at a relative velocity of u.

So:

$$\frac{d\phi}{dt} = \frac{1}{c}\gamma^2(u)\frac{du}{dt}$$

But from the length contraction formula:

$$\frac{dt'}{dt} = \frac{\gamma(u')}{\gamma(u)}$$

Therefore, substituting these two equations in (1):

$$\frac{1}{c}\gamma^2(u)\frac{du}{dt} = \frac{d\epsilon}{dt'}\frac{\gamma(u')}{\gamma(u)}$$

Applying the differential of arctanh as before to determine $\frac{d\epsilon}{dt'}$:

$$\gamma^3(u')\frac{du'}{dt'} = \gamma^3(u)\frac{du}{dt}$$

This is a different result from the Newtonian formula in which $du / dt = du' / dt'$. The **proper acceleration**, α is defined as the acceleration of an object in its rest frame. It is the instantaneous change in velocity for an observer for whom $u' = 0$ and $\alpha = du' / dt'$. In these circumstances:

$$\alpha = \gamma^3(u)\frac{du}{dt}$$

Mathematical Appendix

Mathematics of the Lorentz Transformation Equations

Consider two observers *O* and O', moving at velocity *v* relative to each other, who observe the same event such as a flash of light. How will the coordinates recorded by the two observers be interrelated?

These can be derived using linear algebra on the basis of the postulates of relativity and an extra homogeneity and isotropy assumption.

The homogeneity and isotropy assumption: space is uniform and homogenous in all directions. If this were not the case then when comparing lengths between coordinate systems the lengths would depend upon the position of the measurement. For instance, if $x' = ax^2$ the distance between two points would depend upon position.

The linear equations relating coordinates in the primed and unprimed frames are:

$x' = a_{11}x + a_{12}y + a_{13}z + a_{14}t$

$y' = a_{21}x + a_{22}y + a_{23}z + a_{24}t$

$z' = a_{31}x + a_{32}y + a_{33}z + a_{34}t$

$t' = a_{41}x + a_{42}y + a_{43}z + a_{44}t$

There is no relative motion in the y or z directions so, according to the 'relativity' postulate:

$z' = z$

$y' = y$

Hence: $a_{22} = 1$

$a_{33} = 1$

and: $a_{21} = 0$

$a_{23} = 0$

$a_{24} = 0$

$a_{31} = 0$

$a_{32} = 0$

$a_{34} = 0$

So the following equations remain to be solved:

$x' = a_{11}x + a_{12}y + a_{13}z + a_{14}t$

$t' = a_{41}x + a_{42}y + a_{43}z + a_{44}t$

If space is isotropic (the same in all directions) then the motion of clocks should be independent of the y and z axes (otherwise clocks placed symmetrically around the x-axis would appear to disagree. Hence $a_{42} = a_{43} = 0$ so:

$t' = a_{41}x + a_{44}t$

It is also the case that when $x' = 0$ then $x = -vt$. So:

$0 = a_{11}vt + a_{12}y + a_{13}z + a_{14}t$

and

$-a_{11}vt = a_{12}y + a_{13}z + a_{14}t$

Given that the equations are linear then $a_{12}y + a_{13}z = 0$ and:

$-a_{11}vt = a_{14}t$

and

$-a_{11}v = a_{14}$

Therefore the correct transformation equation for $x^{'}$ is:

$x^{'} = a_{11}(x - vt)$

The analysis to date gives the following equations:

$x^{'} = a_{11}(x - vt)$

$y^{'} = y$

$z^{'} = z$

$t^{'} = a_{41}x + a_{44}t$

The event is a flash of light that expands as a sphere with the following equations in each coordinate system, assuming that the speed of light is constant:

$x^2 + y^2 + z^2 = c^2t^2$

$x^{'2} + y^{'2} + z^{'2} = c^2t^{'2}$

Substituting the coordinate transformation equations into $x^{'2} + y^{'2} + z^{'2} = c^2t^{'2}$ gives:

$a11^2(x - vt)^2 + y^2 + z^2 = c^2(a_{41}x + a_{44}t)^2$

rearranging:

$$(a11^2 - c^2a_{41}^2)x^2 + y^2 + z^2 - 2(va_{11}^2 + c^2a_{41}a_{44})xt = (c^2a_{44}^2 - v^2a_{11}^2)t^2$$

This is equivalent to: $x^2 + y^2 + z^2 = c^2t^2$

So: $c^2a_{44}^2 - v^2a_{11}^2 = c^2$

$$a11^2 - c^2a_{41}^2 = 1$$

$$va_{11}^2 + c^2a_{41}a_{44} = 0$$

Solving these 3 simultaneous equations:

$$a_{44} = \frac{1}{\sqrt{(1 - v^2/c^2)}}$$

$$a_{11} = \frac{1}{\sqrt{(1 - v^2/c^2)}}$$

$$a_{41} = -\frac{v/c^2}{\sqrt{(1 - v^2/c^2)}}$$

Substituting these values into:

$x' = a_{11}(x - vt)$

$y' = y$

$z' = z$

$t' = a_{41}x + a_{44}t$

gives:

$$x' = \frac{x - vt}{\sqrt{(1 - v^2/c^2)}}$$

$y' = y$

$z' = z$

$$t' = \frac{t - (v/c^2)x}{\sqrt{(1 - v^2/c^2)}}$$

The inverse transformations are:

$$x = \frac{x' + vt'}{\sqrt{(1 - v^2/c^2)}}$$

$y = y'$

$z = z'$

$$t = \frac{t' + (v/c^2)x'}{\sqrt{(1 - v^2/c^2)}}$$

License

GNU Free Documentation License

Version 1.2, November 2002

```
Copyright (C) 2000,2001,2002  Free Software Foundation, Inc.
51 Franklin St, Fifth Floor, Boston, MA  02110-1301  USA
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.
```

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified

Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties--for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements."

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.