Part V

Introduction to Classical and Quantum Relativity

Curt Wittig

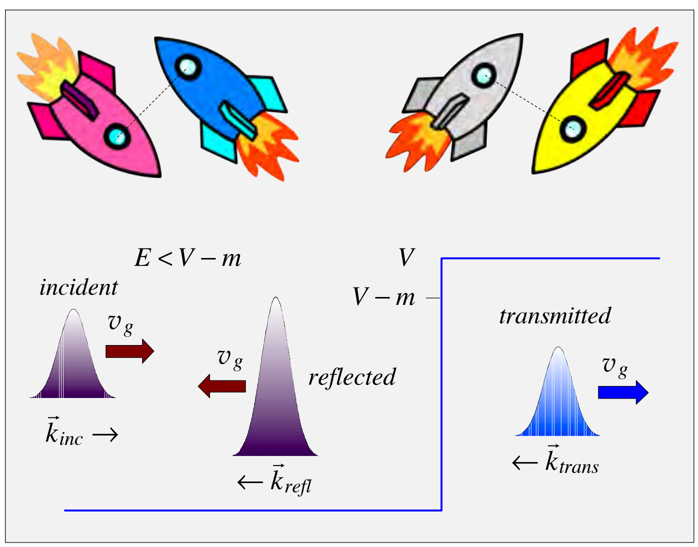
Cover Art by Robert Wittig
Front: Time Will Tell
Rear: Caboose

Preface

This document is a tidied version of notes that were distributed in a class that I taught in the spring semester of 2015. It comprises bits and pieces that had accumulated over the last half-dozen years or so, combined and integrated with material added during the course – roughly equal parts of each. The notes were intended to complement the classroom lectures and textbooks, certainly not replace them.

The course goal was to develop relativistic quantum mechanics systematically, starting with the theory of classical special relativity that debuted at the start of the twentieth century, notably with Einstein's seminal 1905 paper. This "from-the-ground-up" approach is the antithesis of jumping headfirst into the quantum version. I believe that at least as much intuition is gained from working through the basics (by way of examples, exercises, and toy models) as from computer experiments with sophisticated programs. Of course, these approaches complement one another, and their combination is undoubtedly the best way to learn the material. There was no chance that we would get as far as computational relativistic quantum chemistry, given the time and energy constraints. This would be a wonderful follow-on course.

I had not lectured on this material previously, though I had read a fair amount on electrodynamics and relativistic quantum mechanics over the past half-dozen years, working problems and developing a set of typed notes along the way. My expectation at the outset was that the course would be exciting and fun, and this indeed proved to be the case. Besides, a group of students had asked me to offer a course on this material, so there was no chance of going wrong.



The agreement was that we would start at an elementary level with special relativity applied to material objects, work through it with great intensity until exhausted, switch to electrodynamics and do the same, and embark on a brief foray into gauge field theory to unearth the symbiosis between quantum mechanics and electrodynamics that gives us the minimal coupling algorithm. Only then would we enter the realm of relativistic quantum theory, moving deftly from Klein-Gordon to Dirac, while paying due respect to relativistic quantum field theory, though not belaboring it.

This is a lot of material for a one-semester course. And believe it or not the course lasted only a half-semester. It was a hell of a ride, with coffee and brownies and lemon bars and chocolate and the like, group participation, presentations, and lectures three times each week that generally ran over the allotted time of one hour. There is probably a bean-counting bureaucrat somewhere in the university with nothing better to do than complain about the course running past one hour.

After the course ended, I decided to gather together the class notes, clean them up, pass out bound copies to the participants, and mount them on my website. They now constitute Part V of the Notes Project, which by definition is in a state of constant revision. There were a few minor additions, but what you have in this document is, for all practical purposes, identical to the notes that were passed out during the course.

These notes do not hint of a finished product: raw in places, awkward wording, structure issues (mainly order of presentation), errors of all kinds, and myriad other deficiencies. In the spirit of full disclosure, however, this is essentially what was distributed during the course. If and when I teach this material again, there will be a number of improvements. First and foremost, I will never again try to pull it off in a half-semester. A full semester is needed, if not more. The additional time will be devoted to exercises, examples, and student/postdoc presentations. No new material will be added. What is in these notes will be improved and, in fact, streamlined. At least that is my intention.

The participants made the course stimulating and worthwhile: postdocs Shirin Faraji, Samer Gozem, and Thomas Jagau; and graduate students Natalie Orms, Bailey Qin, Arman Sadybekov, Parmeet Nijjar, Dhara Trivedi, Bibek Samanta, and Subhasish Sutradhar. Even Professor Krylov joined us in the beginning. The discussions that arose spontaneously were particularly valuable, especially when there was not a consensus of opinion. Without this special group of participants there is no way that such an enormous amount of challenging material could have been tackled in so short a time. Even this group, however, as unique as they were, would have benefited greatly from the course lasting one-semester rather than a half-semester.

Curt Wittig
June 2015

"Fundamental ideas play the most essential role in forming a physical theory. Books on physics are full of complicated mathematical formulae. But thought and ideas, not formulae, are the beginning of every physical theory. The ideas must later take the mathematical form of a quantitative theory, to make possible comparison with experiment."

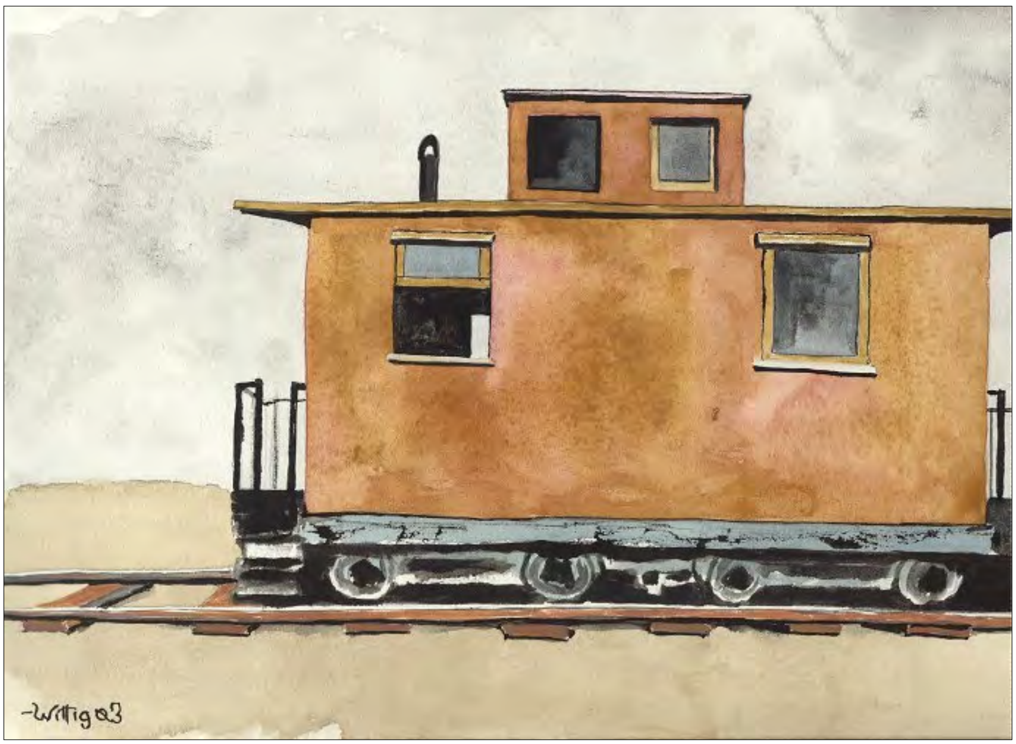
A. Einstein and L. Infeld
The Evolution of Physics (1938)

Contents

Appendices

Chapter 1.

Classical Special Relativity of Material Objects



Caboose
Robert Wittig

One of Einstein's Favorite Gedanken Experiments

A train car moving past a standing observer is struck by two bolts of lightning at different positions: the rear and front of the car. In the inertial frame of the standing observer, three events are spatially distinct but simultaneous: the standing observer facing a moving observer in the center of the car; lightning striking the front of the car; and lightning striking the back of the car. Because the events are along the axis of train movement, their time coordinates project to different time coordinates in the moving train's inertial frame. In the moving train's frame, lightning strikes the front of the car before the two observers face each other.

Contents

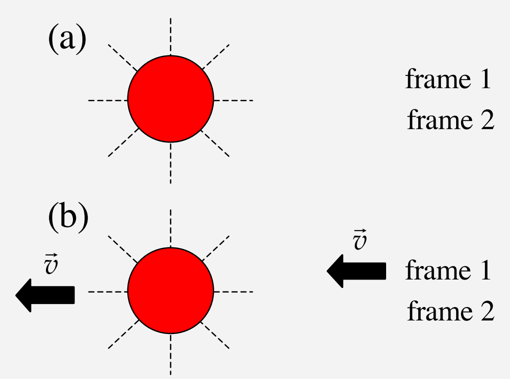
Covariant Representation[‡]

[‡] It is my expectation that these sections will be used as needs arise. They are placed at the end of the chapter, almost like appendices, because they will most likely be referred to a number of times throughout the chapter.

Introductory Comments

The theory of special relativity made its official debut with Albert Einstein's seminal 1905 publication: *On the Electrodynamics of Moving Bodies* [1]. Truth be told, the theory's genesis was convoluted, rife with intrigue, and spread out over decades. By the end of the nineteenth century, it was clear that Newton's laws, Galilean relativity, and Lorentz's ether theory had too many problems to survive. Change was in the air, with scientists like Hendrik Lorentz, Max Planck, Henri Poincaré, and Hermann Minkowski poised for the kill. It is not widely known, but prior to Einstein's paper, Poincaré had already published the mathematical foundation of the Lorentz group, which is a subgroup of the Poincaré group [2-4]. This work is a mainstay of Einstein's theory. It is interesting that it was Poincaré who graciously assigned Lorentz's name to the group. In any event, Einstein's magnificent 1905 paper went well beyond mathematics, and it is the accepted starting point of the theory of special relativity, at least in most scientific quarters.

The theory of special relativity that was set down in Einstein's 1905 paper is premised on two postulates. (1) The laws of physics are the same in all inertial reference frames. Consequently, it is impossible to determine whether a frame is in a state of uniform motion. (2) The speed of light is independent of the motion of its source. This in turn ensures that the speed of light is the same in all inertial reference frames, as illustrated with the sketch on the right.



(a) The light source and observer reference frames are stationary. (b) The light source and frame 1 are made to move toward the left with the same fixed speed. Thus, they are in an inertial reference frame, though a different one than the one from which they started. The observer in frame 2 now detects photons from a moving source. From postulate (2), the speed of light is unaffected. It follows that the speed of light is the same in every inertial frame.

It turns out that the existence of a limiting speed that cannot be exceeded emerges as a direct consequence of the homogeneity and isotropy of empty space.[1] No other assumptions are required. On this basis, the theory can be traced to Bernhard Riemann (1826-1866), one of the most brilliant mathematicians of all time, that is, all time until the present. It was Einstein, though, who attached the physical significance that

[1] The concept of empty space is intriguing, as space and time are meaningless in the absence of objects. As stated by Gottfried Leibniz (1646-1716): "…[space] is that order which renders bodies capable of being situated, and by which they have a situation among themselves when they exist together, as time is that order with respect to their successive position. But if there were no creatures, space and time would only be in the ideas of God." We shall take the term empty space to mean the absence of gravitational forces.

shook the world of physics. It turns out that this limiting speed is that of light in vacuum, again something introduced through Einstein's physical insight.

A misconception that arises every so often when starting out is that special relativity is ill suited to deal with acceleration, as the theory is premised on uniform motion (inertial frames), in which case particles do not undergo acceleration. Were this true, the theory would be pretty useless. Gravitational forces are ignored, but not forces such as those of electrodynamics. The reason gravity is ignored is that the spacetime of special relativity is flat, whereas in general relativity spacetime is curved. In fact, there are straightforward, albeit nontrivial, ways to deal with gravity in special relativity, though we shall not go there. Acceleration in the flat spacetime of special relativity, though straightforward, must be dealt with carefully. This will be discussed at some length.

Spacetime coordinates are established such that non-interacting particles that are at rest remain at rest, while those undergoing uniform motion continue doing so. These are the inertial frames. In the simplest of terms, inertial reference frames are ones that move with constant velocity with respect to one another.[2] Once the inertial frames are in place, we are free to consider the dynamical processes of interest. If we wish to see how things appear from the perspective of a body undergoing acceleration, say through electromagnetic interaction with another object, we use a Lorentz transformation to switch frames. What we refer to as being inertial frames are, in fact, *locally* inertial frames. They satisfy the criteria for being inertial over some (often modest, even infinitesimal) spacetime interval. We can therefore transform from one frame to the next and keep doing this as a particle or object follows its world line in spacetime.

An example that makes the point quite well is that of free-fall in a gravitational field, which is examined thoroughly by Taylor and Wheeler [5]. Imagine a space explorer in a rocket ship that enters a gravitational field. The space ship and everything in it is accelerated toward the source of the field. However, the explorer experiences no acceleration relative to the spaceship. If the explorer pushes from a wall to acquire a small velocity, this velocity will persist until contact is made with the opposite wall. The spaceship serves as an inertial frame for the explorer as long as the spaceship is not too large and measured dimensions are not too small. The spaceship then serves as a *locally* inertial frame. Taylor and Wheeler deal at length with acceleration in a gravitational field, and they give examples of how to estimate its effect, for example showing that it is usually small. Keep in mind that on earth we are in a state of constant acceleration, with the earth pulling us toward its middle, and its surface countering this by pushing us in the opposite direction. Yet, special relativity works quite well.

---

[2] A definition of an inertial frame is given by Taylor and Wheeler [5]: "A reference frame is said to be inertial in a certain region of space and time when, throughout that region of spacetime, and within some specified accuracy, every test particle that is initially at rest remains at rest, and every test particle that is initially in motion continues that motion without change in speed or in direction." An example of an inertial frame is the following. Two clocks are synchronized at the same location (point 1). One is then moved to a different location (point 2). A light pulse released from point 1 at t_1 arrives at point 2 when the clock there reads t_2. The light pulse bounces off a mirror at point 2 and goes back to point 1, where the clock reads t_1' upon the pulse's arrival. If t_2 is one half times the time-difference, $t_1' - t_1$, the system is said to be inertial.

We deal routinely with something of a similar nature in molecular spectroscopy. When viewed from afar, a rotating molecule is in a state of constant acceleration. Solving the mathematical problem from this reference frame would prove formidable. Upon transforming to the rotating frame, however, the math settles down and some new terms appear. These are apparent forces: Coriolis and centripetal. They arise because of the frame transformation.

The fact that the laws of physics are the same in all inertial reference frames does not mean that quantities measured in these frames have the same values. Some do and some do not. For example, we shall see that an object's time and spatial coordinates measured in one inertial frame do not have the same values as the time and spatial coordinates of the same object measured relative to a different inertial frame. The same applies to other "four-vectors." On the other hand, scalar quantities retain their values in all inertial frames. We shall see that this requires clear definition of the term scalar. For example, electric charge density is a scalar in the non-relativistic theory, whereas it is a component of a four-vector in the relativistic theory. A given *amount* of charge is preserved in going between frames, but not the charge's density.

More mathematically elegant and sophisticated formulations of the theory have their place. However, in the beginning we shall forego these in favor of a pedestrian but accessible approach. There is no limit to the amount of math that can be brought to bear, for example, were we to venture beyond special relativity to general relativity. Some of this can prove helpful, but only after the basics are in place. In light of this, after the essential concepts and results have settled, a complementary development will be carried out using a more succinct approach. Certain aspects of this, such as covariant and contravariant notation, spacetime metric, summation convention, and so on are extremely useful. Consequently, these will be introduced along the way. Indeed, you are encouraged to scan Sections 4-6 after completing Section 2, if not before. Sections 4-6 should be treated like super-appendices in the sense that they are enlisted whenever the need arises.

Special relativity is about relationships between space and time and the fields and material objects that space and time serve to locate. It plays out on the platform of intertwined space and time called spacetime, which is central to observations in and between inertial reference frames. The physical science of special relativity, including the quantum version, is rich in nuance, phenomena, and applications. For example, neither the physical properties nor chemistry of heavy elements can be understood without relativity. Indeed, electrons in such systems routinely achieve speeds close to c, and in many cases spin-orbit interaction affects systems profoundly. Numerous consequences of the theory appear counterintuitive at first, which makes the subject even more interesting. For example, the simultaneity that we accept without question in everyday life is a casualty, as noted on the cover page with one of Einstein's favorite gedanken (thought) experiments.

In putting together the notes for this chapter, I have relied a great deal on textbooks, especially the classic ones by Wolfgang Rindler (Relativity: Special, General, and Cosmological) [6], and by Edwin Taylor and John Archibald Wheeler (Spacetime Physics) [5]. Other textbooks [7-12] have also proven useful, each in its special way. Undoubtedly there are many others that might prove useful. If you come across one that is to your liking, let me know. It will be ordered and placed in the course library.

Wolfgang Rindler (1924 –)

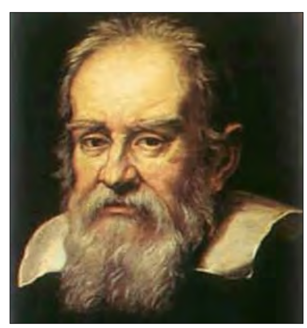
John Wheeler (1911 – 2008)

Edwin Taylor

Rindler was educated in England, having arrived there in 1938 at age fourteen as part of the Kindertransport that rescued Jewish children from the Nazis. He has worked in the United States since 1956 and is presently at the University of Texas at Austin. He introduced the term "event horizon." This arises in general relativity, usually in the context of black holes. We will touch on it lightly. John Archibald Wheeler, in addition to being a brilliant scientist, was a dedicated teacher who was fortunate to mentor a number of exceptional graduate students, not the least of which was Richard Feynman. I was not able to unearth much about Edwin Taylor. What I did find is that he served as editor of the American Journal of Physics, and he was held in quite high regard as a teacher. For example, he received the 1998 Oersted Medal for his teaching prowess.

Copies of these and a number of other books on (mainly special) relativity have been placed in the course's mini-library in room 401. They are also listed in the Bibliography and References, including comments regarding content, level, and so on. Before starting in earnest with special relativity, a few comments are in order regarding its predecessor. After all, deep thought about reference frames and transformations between them was around for a long time before the start of the twentieth century.

Galilean Relativity



Galileo Galilei (1564-1642)

The legendary Italian polymath (Renaissance Man), Galileo Galilei, was the first person to introduce a scientific model for transformations between inertial reference frames. Galileo's theory of relativity, which predates the laws of physics set down by Isaac Newton, is the picture of simplicity. It survived for roughly 300 years, succumbing finally to Einstein's theory. Its competition began to gather momentum in the mid nineteenth century, and by the dawn of the twentieth century its demise was inevitable. It was Einstein's 1905 paper, though, that sounded the death knell for Galilean relativity.

In what follows, you will be presented with a distilled version of the reasoning behind Galileo's relativity, ending with the transformation that bears his name.[3] This material is important: the context in which Einstein and others put forth the moves that resulted in the theory of special relativity.

[3] According to Rafael Ferraro [7], this transformation was assigned the label "Galilean transformation" in 1909 by P. Frank. I have been unsuccessful at unearthing anything about P. Frank.

Galileo came to the conclusion that it is not possible to establish whether a frame is in a state of uniform motion: "…any velocity once imparted to a moving body will be rigidly maintained as long as the external causes of acceleration or retardation are removed, a condition which is found only on horizontal planes; for in the case of planes which slopes downwards, there is already present a cause of acceleration, while on planes sloping upward there is retardation; from this follows that motion along a horizontal plane is perpetual; for, if the velocity be uniform, it cannot be diminished or slackened, much less destroyed." Galileo argued against the existence of a preferred reference frame (universal rest frame), which put him at odds with the ether theory more than two-and-a-half centuries before it even existed. It also put him at odds with the Catholic Church, for which he came close to being executed.

Newton imposed mathematical structure on Galileo's reasoning. He used the calculus that he and his nemesis, Gottfried Leibniz, had invented to show that acceleration is an invariant, and of course he identified force as mass times acceleration. Newton paid homage to Galileo by naming the transformation equations after him.

Isaac Newton
(1642-1727)

To begin, we shall consider two Cartesian reference frames and distinguish between them by using unprimed and primed labels. A point has coordinates (x, y, z) in the unprimed frame and coordinates (x', y', z') in the primed frame. The coordinates are related to one another through the frame transformation

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = R \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} + T \,. \tag{1}$$

The matrix R rotates the primed frame, and the column vector T moves it to another location [9]. The operations R and T are, in general, time dependent. Let us now differentiate the above equation twice with respect to time, thereby obtaining the acceleration:

$$\begin{pmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{pmatrix} = R \begin{pmatrix} \ddot{x}' \\ \ddot{y}' \\ \ddot{z}' \end{pmatrix} + 2\dot{R} \begin{pmatrix} \dot{x}' \\ \dot{y}' \\ \dot{z}' \end{pmatrix} + \ddot{R} \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} + \ddot{T} \,. \tag{2}$$

Acceleration is a physical vector that does not change in going between inertial reference frames. Note that the force a person experiences does not depend on the inertial frame from which the person is observed. For example, a standing person who measures a passing car's acceleration gets the same value, within experimental error, as is measured by a person in a car that is moving at a constant speed. There are higher derivatives as well: $\dddot{x}$, $\ddddot{x}$, etc, but here we need only consider $\ddot{x}$.

The acceleration's components (which are represented as entries in a column vector) transform according to eqn (1), whereas the vector itself is the same in all inertial frames. Velocity changes from one inertial frame to another, but not acceleration. It is important to distinguish physical vectors, the kind we represent using arrows, from column vectors. The latter are a representation of the vector in a basis. In the present case, the axes of the primed and unprimed frames constitute bases.

Referring to eqn (2), we see that the acceleration is the same in the two frames if and only if $\ddot{T} = \dot{R} = 0$. In other words, it is acceptable for the frames to move with constant velocity with respect to one another ($\dot{T} \neq 0$), but they must not rotate. When these conditions are satisfied, we have inertial frames. Equation (2) also verifies that acceleration is invariant with respect to transformation between inertial frames.

The Galilean transformation also includes time, so $(x,y,z) \to (x',y',z')$ is extended to $(x,y,z,t) \to (x',y',z',t')$. Of course, t and t' can only differ by a constant, whereas $\dot{T} \neq 0$ means that the frames can have a relative velocity $\vec{v}$ as well as a constant offset. The offset can indicate how far apart the frames are at a given time, say at $t = 0$. With the above considerations in hand, we now express the Galilean transformation in matrix form:

$$\begin{pmatrix} t \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ v_x & R_{11} & R_{12} & R_{13} \\ v_y & R_{21} & R_{22} & R_{23} \\ v_z & R_{31} & R_{32} & R_{33} \end{pmatrix} \begin{pmatrix} t' \\ x' \\ y' \\ z' \end{pmatrix} + \begin{pmatrix} C_t \\ C_x \\ C_y \\ C_z \end{pmatrix}. \tag{3}$$

The column vector C accounts for (constant) offsets in time and space, the 3×3 orthogonal matrix R is a frame rotational displacement (having no time dependence), and the components v_i of the relative velocity $\vec{v}$ are self-explanatory. No issue arises with simultaneity, and the speed of light is nowhere to be seen, meaning that all speeds are possible. Notice that the replacements $t \to ct$, $t' \to ct'$, and $v_i \to v_i / c$ put time and space on more equal footing, while leaving the four component equations in eqn (3) unaffected. The bottom line is that everything behaves sensibly according to Newtonian mechanics. Later we will see that in the non-relativistic limit, the Lorentz transformation – the cornerstone of special relativity – reverts to eqn (3).

Without rotational displacement between frames [*i.e.*, R = diag (1,1,1)], what remains is referred to as a Galilean boost. For example, with $v_y = v_z = 0$, eqn (3) gives $t = t'$ and $x = x' + v_x t + C_x$. The constant C_t is zero when t and t' share a common starting point, say $t = t' = 0$. Likewise, the constants C_i are each zero when the frames share a common origin, say at $t = t' = 0$. And so on.

With the Galilean transformation as a backdrop, let us now begin.

1. Lorentz Transformation: The Boost

The transformation that relates measurements of space and time in one inertial frame to measurements of space and time in another inertial frame is derived in this section. This transformation is referred to as a Lorentz boost, relativistic boost, or simply a boost. It is the cornerstone of the theory of classical special relativity. Chapter 1 deals almost exclusively with material objects. Photons are enlisted on occasion, mainly as measuring devices, but Maxwell's equations never surface. Extension to electrodynamics is carried out in Chapter 2.

The transformations of interest vary continuously about the identity: 3D rotation and the relativistic boosts derived in the present section. Together these constitute the "Proper Orthochronous Homogeneous Lorentz Group," which shall be referred hereafter to as the Lorentz Group. The discrete symmetries: parity and time reversal, though unimportant in the present context, enter in the quantum version discussed in Chapter 4.

Henri Poincaré was a mathematician who also taught electromagnetism. He derived a more general result that includes translational displacement: the Poincaré group. This displacement is counterpart to the displacement *C* in the Galilean transformation. The Poincaré group is close to the Lorentz group (which was also derived by Poincaré), differing only in the displacement, and it is important in other theories. We shall work exclusively with the Lorentz Group.

Contributors to special relativity, from the left: George Fitzgerald, Henri Poincaré, Hermann Minkowski, Albert Einstein, and Hendrik Antoon Lorentz.

George Fitzgerald and Hendrik Antoon Lorentz in the late nineteenth century derived what came to be known as the Lorentz transformation. A number of scientists participated in the development of the theory of special relativity, which has a rich, amusing history. At the time, the Lorentz transformation was considered revolutionary, as its implementation in physics had counterintuitive consequences: shortening of a dimension of a moving object (referred to as Lorentz contraction, Lorentz-Fitzgerald contraction, length contraction, or simply contraction), time dilation, and loss of simultaneity. The majority of scientists, of course, stuck with the Galilean transformation. Truth be told, the majority of scientists failed to appreciate special relativity until the 1920's, if not later. After all, Einstein's 1921 Nobel Prize was for the photoelectric effect, not relativity, as the latter was still considered controversial in some quarters. Habits of early satisfaction die hard.

The Lorentz transformation was used with the ether theory prior to its use by Henri Poincaré and Einstein in special relativity. Derivations that yield the Lorentz boost can be found in many texts. Two are given here. One is succinct, which earns it a high mark, whereas the other is lengthier but perhaps easier to follow.

1.1. Frames

Figure 1.1 indicates two inertial Cartesian reference frames: a primed one with origin O', and an unprimed one with origin O. It is assumed that the unprimed frame is stationary, whereas the primed frame moves to the right with constant velocity $\vec{v} = v\hat{x}$. There is no need to keep one frame stationary. However, this is traditional and easy to visualize. Time is recorded in the two frames such that at $t = 0$ and $t' = 0$ the origins O and O' coincide. The horizontal axis in the yellow box has been displaced vertically from the lower horizontal axis for viewing convenience. Likewise, the assumption that t and t' are each equal to zero when O and O' coincide is for convenience. For $t > 0$, an observer in the unprimed frame can measure the location of O' as well as the locations of objects that are either moving or stationary in O'. At time t, the origin O' is located at $x = vt$. The issue of which frame is moving and which is stationary is a matter of perspective. Were O' taken as stationary, O would be moving toward the left with velocity $\vec{v} = -v\hat{x}$.

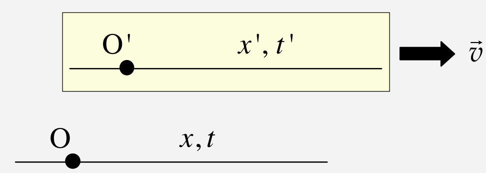


Figure 1.1. Two inertial reference frames: The one in the box, whose origin is O', moves with constant velocity $\vec{v}$ relative to the stationary one, whose origin is O. Axes x and x' share a common horizontal line. They are displaced vertically from one another to facilitate visualization.

Now consider a case in which a pulse of light is launched in the O' frame. Figure 1.2(a) depicts a light pulse that commences travel at $t = 0$ from M_1 to M_2. These mirrors are stationary in O'. An observer in O' records the fact that the pulse bounces back and forth between the mirrors.

The O' frame moves to the right with constant velocity $\vec{v}$ relative to O. An observer in O also is able to record the light pulse. To this observer the mirrors are moving to the right with velocity $\vec{v}$. Thus, the observer in O finds that the light pulse moves in a straight line that makes an angle with respect to the vertical, as indicated in Fig. 1.2(b) with red "snapshots." Namely, as the light pulse goes from M_1 to M_2, it travels downward and to the right when measured from O. To make a nice figure, a large velocity was used, that is, $|\vec{v}|$ is a significant fraction of c in Fig. 1.2(b). I am sidestepping here the issue of exactly how these observations are carried out. This is explained in the next subsection.

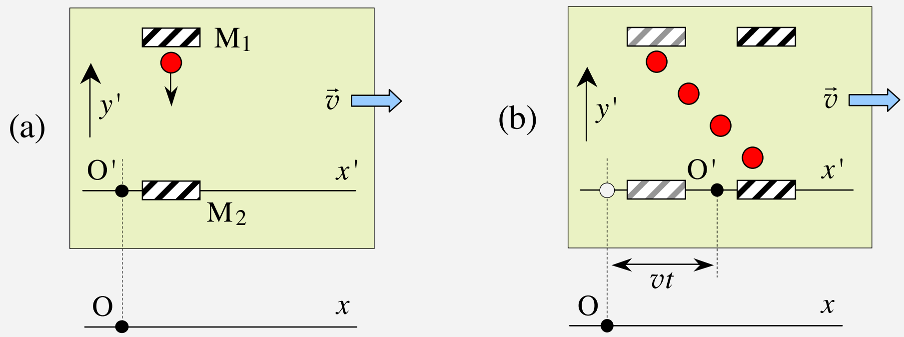


Figure 1.2. (a) Axes x and x' share a common horizontal line. They are displaced vertically from one another to facilitate visualization. A light pulse travels between mirrors M_1 and M_2. The primed system (boxed) moves to the right with velocity $\vec{v}$ relative to O. To an observer in O', the light pulse move vertically downward, as indicated with the vertical arrow. In (b), the primed system has moved to the right a distance vt. The original position of the mirrors is denoted by gray slanted lines, their position at $x = vt$ by black slanted lines. An observer in the O frame finds that throughout the mirror's transit the light pulse follows a straight-line path whose length exceeds the mirror separation, l.

The mirror separation in O' is $l = ct'$, where t' is the time measured in O' that it takes the light pulse to travel between mirrors. Again, it is assumed that the origins, O and O', coincide at $t = t' = 0$. For an observer in O, the distance traveled by a light pulse that goes from M_1 to M_2 is larger than the mirror separation. Specifically, because the mirrors move to the right with constant speed during the transit of the light pulse, the light path is tilted toward the right. To take this into consideration, the distance traveled by the light pulse, as viewed from O, is assigned the value γl. The dimensionless parameter γ is a scale factor that accounts for the increased length of the path traveled by the light. It always exceeds unity.

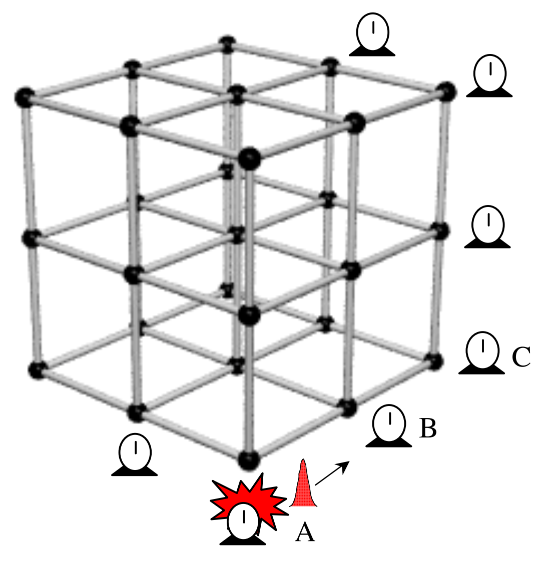
Because $\gamma l = c\gamma t' = ct$, it follows that $t = \gamma t'$. Notice that we are using the fact that the speed of light is the same in both reference frames. This ensures that an observer in O finds that it takes a longer time for the light pulse to travel between mirrors than the mirror-to-mirror transit time recorded in the O' frame. Because γ always exceeds unity, the observer in O concludes that time passes more slowly in the moving reference frame. The clock in O' is slow and may need to be fixed. For example, maybe the observer in O' found that it took 10 ns for the light to go from M_1 to M_2, whereas the observer in O found that it took 11 ns. We shall see that an important consequence is that events that take place simultaneously in non-relativistic mechanics ($c = \infty$) do not, in general, take place simultaneously in relativistic mechanics.

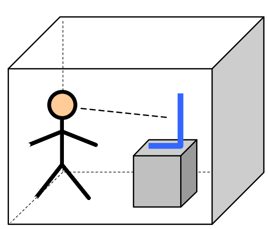
1.2. An Observer and Her Lattice of Clocks

This is a good time to discuss the meaning of the ubiquitous "observer" of special relativity. Two kinds of observation shall be considered. In the first, the observer does not depend on her visual acuity. Rather, she makes use of a lattice of points in 3D space with a separate clock attached to each point, as indicated in Fig. 1.3. If a particle appears at one of the lattice points, the coordinates of this point and the time on its clock are recorded. For example, this information is stored in a computer for later retrieval. The particle then appears at another location at a later time. Again, the point and time are recorded and stored for later retrieval. And so on as the particle moves through the lattice.

This method of observation does not require that the observer watch visually what is going on. Instead, she relies on the "lattice of clocks" that spans her inertial frame. Space and time resolution are determined by the lattice spacing, which can assume any value, as this is a gedanken experiment. Thus, you can think of the particle's movement as following a smooth trajectory, with space and time resolution (of the observer's choosing) provided by the lattice of clocks.

Figure 1.3. A cubic lattice of points (nearest neighbor spacing of 1 meter) fills an inertial frame. Each point has a clock attached to it. When clock A reads $t = 0$, a short duration light pulse is released. If it arrives at clock B when clock B reads $t = c^{-1}$, clocks A and B are synchronized. And so on. The observer peruses the record of events, for example, as a particle followed a path through the inertial frame, from the comfort of her office.





For the above method of observation to work, the clocks must be synchronized with respect to one another. There are a number of ways to achieve this. Here we shall use perhaps the simplest. Referring to Fig. 1.3, a lattice of points extends throughout the observer's inertial rest frame. The points are separated from their nearest neighbors by 1 meter. Consider clocks A, B, and C. At the instant clock A registers $t = 0$, it releases a short-duration light pulse (say 20 fs) that is directed toward clocks B and C. When the light pulse arrives at clock B, the reading of this clock should be $t = c^{-1}$ (roughly 3.3 ns). Likewise, when the light pulse arrives at clock C its reading should be $t = 2c^{-1}$. If the entire lattice of clocks follows this recipe, the clocks are taken as synchronized.

When the pulse arrives at clock B, were an observer located at clock B to look visually at clock A she would see it registering $t = 0$ because of the 3.3 ns delay time, namely, the travel time for light going from A to B. The lattice of clocks method of observation is used in the construction of the theory.

Now consider a method of observation in which the observer records what she sees visually. This differs from the lattice of clocks because the light that reaches the observer's eye at a given instant originates from different locations. You can replace the observer's eye with a camera having a fast shutter.

Imagine an object in motion relative to the observer. Different parts of the object must send out photons at different times if they are to arrive at the observer's eye at a given moment of "observation." In the non-relativistic regime, this is unimportant, and the observer sees the object as it exists in its rest frame. In the relativistic regime interesting effects can arise, as illustrated in examples presented later that highlight the differences and interplay between the lattice of clocks and visual observation methods.

Aside from pointing out interesting features that arise in cases of visual observation, we shall deal almost exclusively with observations carried out using lattices of clocks. Whenever reference is made to an observer, it is understood that this means that a lattice of clocks records positions and times in the reference frame wherein the observation is made. Do not get me wrong: Real observations of relativistic phenomena in deep space using the best telescopes ever developed are visual observations that need to be interpreted, and this will be discussed. For gedanken experiments, though, lattices of clocks are preferred. Let us now return to what is referred to as the relativistic boost.

Exercise: An observer has a lattice of clocks that has flashlamps attached to each lattice point. Each of these flashlamps radiates isotropically. Let us say that the nearest-neighbor distance between lattice points is 1 cm, whereas the flashlamps produce pulses of duration 10 μs. At a prescribed time, all of the lamps in the lattice are fired – simultaneously in the observer's rest frame. What does the observer see visually?

1.3. Relativistic Boost

Referring to Fig. 1.2, the origins O and O' are taken as coinciding at $t = t' = 0$. To avoid needless repetition, let us agree that the labels O and O' can refer to either the respective origins or their associated reference frames, with distinction drawn from context. At time t, an observer in O, using her lattice of clocks, finds that the distance between O and O' is vt. This can be written: $vt = v\gamma t' = v\gamma l / c = \beta\gamma l$, where $\beta = v / c$. As discussed in the previous subsection, the observer in O finds that the light pulse travels a distance γl as it goes from one mirror to the other, whereas the observer in O' finds that the light pulse travels a distance l as it goes from one mirror to the other. The observers in O and O' agree that the mirror separation is l.

To obtain an expression for the scale factor γ, these facts are used to construct the right triangle shown in Fig. 1.4. As mentioned above, measurements carried out in the O reference frame indicate that the horizontal distance is $vt = \beta\gamma l$, the vertical distance is l, and the hypotenuse is γl. This gives: $\gamma^2 = 1 + \beta^2\gamma^2$, which is written

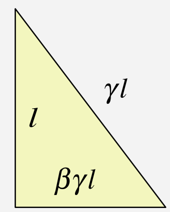


Figure 1.4. Referring to Fig. 1.2 and the accompanying text, the relevant distances are assigned to the sides of a right triangle.

$$\gamma^2 = \frac{1}{1-\beta^2} , \tag{1.1}$$

where $\beta = v / c$.

It is now assumed that the properties of empty space are the same everywhere (space homogeneity), and that these properties are the same in all directions (isotropy). Likewise, it is assumed that carrying out an experiment at one time gives the same results as carrying it out at another time (temporal homogeneity). This ensures that transformations between spacetime coordinates measured in different coordinate frames obey linear relationships.[4] Because the transformation between the primed and unprimed systems is linear, x and t can be expressed in terms of x' and t'. The algebra that is used to obtain expressions for x and t in terms of x' and t' is given in the box below. The result is

---

[4] Suppose that x is a function of x' and t', *i.e.*, $x = f(x', t')$. In this case, we have

$$dx = \frac{\partial f(x', t')}{\partial x'} dx' + \frac{\partial f(x', t')}{\partial t'} dt' .$$

The homogeneity requirement demands that the partial derivatives are constants. This ensures that a change dx is the same at all spacetime locations. Inclusion of y and z coordinates is trivial. Thus, the transformation is linear. The isotropy requirement is used later. Namely, when taking O' as stationary, O travels with a velocity $\vec{v} = -v\hat{x}$.

$$\begin{pmatrix} ct \\ x \end{pmatrix} = \begin{pmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} ct' \\ x' \end{pmatrix}. \tag{1.2}$$

This describes the relativistic boost. Note that the other distances are unchanged: $y = y'$ and $z = z'$. To invert eqn (1.2) into an expression for the (ct', x') column vector in terms of a 2×2 matrix times the (ct, x) column vector simply exchange primed and unprimed, and let $\beta \to -\beta$, which reverses the direction of the velocity:

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix}. \tag{1.3}$$

There is no significant difference between eqns (1.2) and (1.3), simply a matter of perspective. Equation (1.2) expresses the fact that the primed frame moves in the $+x$ direction relative to the unprimed frame. Equation (1.3) expresses the fact that the unprimed frame moves in the $-x$ direction relative to the primed frame.

2D Transformation: *x* and *t*

The linear transformation between the primed and unprimed systems is written

$$ct = C_1 ct' + C_2 x' \tag{i}$$

$$x = C_3 x' + C_4 ct'. \tag{ii}$$

The C_i are constants. Referring to Figs. 1.2 and 1.3 and the accompanying discussions, when $x = vt$ and $x' = 0$, t is equal to $\gamma t'$. Using these conditions with eqn (ii) yields $C_4 = \beta\gamma$. Likewise, using $t = \gamma t'$ and $x' = 0$ with eqn (i) yields $C_1 = \gamma$.

In exchanging the roles of the primed and unprimed systems, when $x' = -vt'$ and $x = 0$, t' is equal to γt. Introducing these conditions and $C_4 = \beta\gamma$ into eqn (ii) yields $C_3 = \gamma$, while using these conditions with eqn (i) yields $C_2 = \beta\gamma$. In summary, the values: $C_1 = C_3 = \gamma$ and $C_2 = C_4 = \beta\gamma$, yield eqn (1.2).

This derivation is compact. Though brevity has its merits, the step-by-step derivation below may provide more insight. In any event, the pair of them surely gets the message across. Incidentally, Fitzgerald and Lorentz spent a great deal of time deriving the relationship given by eqn (1.2).

1.4. Homogeneity, Isotropy, and Lorentz Boost

We again start with the linear transformation between stationary and moving inertial reference frames that follows from the homogeneity and isotropy of empty space. This transformation is given below in its most general form, that is, before taking any symmetry arguments into consideration. This time we shall forego appending c to t and t' until the end.

$$\begin{pmatrix} x' \\ t' \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix} \xrightarrow{\text{invert}} \begin{pmatrix} x \\ t \end{pmatrix} = \frac{1}{AD - BC} \begin{pmatrix} D & -B \\ -C & A \end{pmatrix} \begin{pmatrix} x' \\ t' \end{pmatrix}. \tag{1.4}$$

Two of the four constants are obtained by evaluating expressions at $x' = 0$ and $x = 0$. Using $x' = 0$ in the left equation yields $x / t = -B / A = v$, where v is the speed with which O' travels in the $+x$ direction. Using $x = 0$ in the right equation yields $x'/t' = B / D = -v$. In this case, use is made of the fact that O moves in the $-x$ direction relative to a stationary O' frame. Combining $B = -vA$ and $B = -vD$ gives A = D.

It will prove convenient later to have the following labels: $A = \gamma$ and $C = \gamma K$, so these are introduced here. The parameters γ and K do not depend on location or time. However, they are free to depend on v, and this will prove to be the case. Keep in mind that at this point γ and K are labels and nothing more. In other words, nothing has been introduced or assumed about the speed of light, either its value or whether it constitutes a fundamental speed limit in the universe. Light is not present in eqn (1.4). The only assumptions that have been introduced are the homogeneity and isotropy of empty space, which enabled us to write the general linear transformation.

Applying the above relations to eqn (1.4) yields

$$\begin{pmatrix} x' \\ t' \end{pmatrix} = \gamma \begin{pmatrix} 1 & -v \\ K & 1 \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix} \xrightarrow{\text{invert}} \begin{pmatrix} x \\ t \end{pmatrix} = \frac{1}{\gamma(1 + vK)} \begin{pmatrix} 1 & v \\ -K & 1 \end{pmatrix} \begin{pmatrix} x' \\ t' \end{pmatrix}. \tag{1.5}$$

Let us now use the fact that the expressions for x and t also can be obtained from those for x' and t' by changing v to $-v$, and allowing γ and K to become γ' and K'. Thus, the equation for the x', t' column vector becomes

$$\begin{pmatrix} x' \\ t' \end{pmatrix} = \gamma \begin{pmatrix} 1 & -v \\ K & 1 \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix} \xrightarrow{v \to -v,\, \gamma \to \gamma',\, K \to K'} \begin{pmatrix} x \\ t \end{pmatrix} = \gamma' \begin{pmatrix} 1 & v \\ K' & 1 \end{pmatrix} \begin{pmatrix} x' \\ t' \end{pmatrix}. \tag{1.6}$$

where K' and γ' are the primed-to-unprimed counterparts to K and γ.

The parameters γ and γ' must each be positive, and from the arbitrariness of choosing which frame is stationary, it follows that neither γ nor γ' can depend on the orientation of the velocity. This is another statement of the assumed isotropy of space, which was used in letting v go to $-v$. It follows that $\gamma = \gamma'$. Comparing eqns (1.5) and (1.6), we also see that $K = -K'$.

Referring to eqn (1.6), we now transform x and t first to x' and t' [left side of eqn (1.6)], and then transform x' and t' back to x and t [right side of eqn (1.6)]. This multiplication, by definition, yields the unit matrix:

$$\mathbf{1} = \gamma^2 \begin{pmatrix} 1 & -v \\ K & 1 \end{pmatrix} \begin{pmatrix} 1 & v \\ -K & 1 \end{pmatrix} \tag{1.7}$$

$$= \gamma^2 \begin{pmatrix} 1+vK & 0 \\ 0 & 1+vK \end{pmatrix}. \tag{1.8}$$

Thus, we have arrived at an expression for γ^2:

$$\gamma^2 = \frac{1}{1+vK}. \tag{1.9}$$

Note that this can also be written straightaway by comparing eqns (1.5) and (1.6).

The parameter K has dimension velocity^{-1}, and it must be proportional to v to ensure that γ does not depend on the choice of which reference frame is taken as stationary. On this basis, we write: $K = -v/\mathsf{v}^2$, where v^2 is a parameter that is yet to be specified.[5] The minus sign is necessary to ensure that transformations such as eqn (1.6) are sensible. This might not be immediately obvious, but if you play with eqn (1.6), you will find that it is true. Were we to use a plus sign here, we would discover later that this choice would not work, and we would have to exchange the plus sign for a minus sign.

A profound result has been obtained on the basis of the homogeneity and isotropy of free space, namely, the expression:

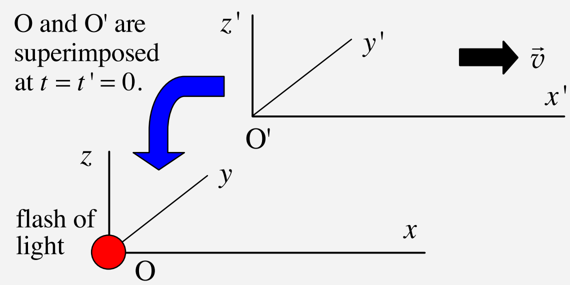
$$\gamma^2 = \frac{1}{1-(v/\mathsf{v})^2}. \tag{1.10}$$

When this is introduced into the transformation equations, it ensures that the speed v must be less than v. This is amazing. The existence of a universal speed limit has been established without recourse to electromagnetic radiation.

[5] Higher powers of v (3, 5…) are ruled out because the transformation equations get crazy.

1.5. Burst of Light

The parameter v^2 is now evaluated. It can be shown that it is a constant. This takes a fair amount of tedious algebra, which I will forego at the moment, and most likely work out later and put in a box or a large footnote. Again, the amazing thing is that we have come so far on the basis of so elementary an ansatz as the homogeneity and isotropy of empty space. We see that space and time are intertwined, and a limiting speed has emerged – a speed that cannot be surpassed by any material object.



The O and O' frames are superimposed when an isotropic burst of light goes off at their origins. Observers in each frame record the advancement of a spherical wave front.

Referring to the above sketch, a short burst of light is emitted isotropically (red circle) from a point source. The origins of the reference frames O and O' are assumed to coincide with the point source at exactly the time when the burst of light is emitted. As usual, O is stationary and O' moves in the $+\,x$ direction with speed v. The expanding spherical wave front recorded by an observer in O satisfies

$$(ct)^2 - x^2 - y^2 - z^2 = 0 \ . \tag{1.11}$$

Likewise, in O' there is a spherical wave front that satisfies

$$(ct')^2 - (x')^2 - (y')^2 - (z')^2 = 0 \, . \tag{1.12}$$

If Lorentz transformation is applied to the latter expression, it will recover the unprimed expression in eqn (1.11), subject to a condition on v^2. This is how v^2 is to be evaluated. Upon carrying out the algebra (box below), the fact that v^2 must be equal to c^2 falls neatly into place, yielding eqn (1.1):

$$\gamma^2 = \frac{1}{1-\beta^2} \ . \tag{1.1}$$

Referring to eqn (1.5), we write

$$
\begin{aligned}
(x')^2 &= \gamma^2 (x - vt)^2 \\
&= \gamma^2 \left(x^2 - 2xvt + (vt)^2\right) \\
(ct')^2 &= \gamma^2 \left(ct - \frac{cvx}{\mathsf{v}^2}\right)^2 \\
&= \gamma^2 \left((ct)^2 - 2\frac{c^2 xvt}{\mathsf{v}^2} + \left(\frac{cvx}{\mathsf{v}^2}\right)^2\right)
\end{aligned}
$$

The y' and z' parts are uninteresting: $(y')^2 = y^2$ and $(z')^2 = z^2$. Combining the x' and ct' parts yields

$$(ct')^2 - (x')^2 = \gamma^2 (ct)^2 \left(1 - \frac{v^2}{c^2}\right) - \gamma^2 x^2 \left(1 - \left(\frac{cv}{\mathsf{v}^2}\right)^2\right) + 2xvt\left(1 - \frac{c^2}{\mathsf{v}^2}\right).$$

The last term must vanish, and this requires $\mathsf{v}^2 = c^2$. When this is introduced into the above equation, we are left with

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

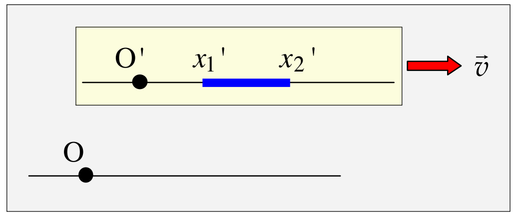
as required. Notice that the choice $K = +v/\mathsf{v}^2$ would not work because it would not be possible to cancel the $2xvt$ term. This confirms the correctness of the earlier choice of $K = -v/\mathsf{v}^2$.

Exercise: Explain why the term $2xvt(1 - c^2/\mathsf{v}^2)$ must vanish.

The above derivation yields the Lorentz boost of special relativity, which is the theory's signature feature. It does so by invoking the assumptions that space is homogeneous and isotropic, the arrow of time, though advancing ever forward, is homogeneous (the same results are obtained at different times), and by demanding that a Lorentz boost is consistent with Einstein's postulate: in effect, that the speed of light in vacuum has the same value in all inertial reference frames. This Einstein postulate is, of course, an experimental fact.

This is a stunning scientific achievement. Spacetime emerges as the fundamental 4D "ruler." Parity and time reversal symmetries are seen to be essentially the same, *i.e.*, inversions in spacetime. These symmetries do not appear in our classical treatment, as they are discrete rather than continuous. However, they are important in relativistic quantum mechanics.

Length Contraction Along the Direction of Motion



A rod that is at rest in the O' frame is aligned with the direction of this frame's motion relative to the stationary O frame, as indicated in the sketch on the right. An observer in O' walks over to the rod with a ruler and records the coordinates x_2' and x_1', thereby obtaining its length: $x_2' - x_1' = l_{O'}$.

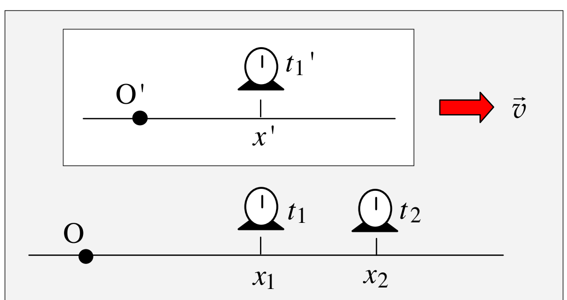
Let us now determine the value that is obtained for the length of the rod when an observer in the stationary O frame measures the rod's length. This length is $x_2 - x_1 = l_O$. Keep in mind that the measurement that is carried out relative to the O frame needs to be made at a fixed time. In other words, x_1 and x_2 need to be measured simultaneously. Lorentz transformation is used to make the comparison:

$$x_2' = \gamma x_2 - v\gamma t \tag{1.13}$$

$$x_1' = \gamma x_1 - v\gamma t \,. \tag{1.14}$$

Subtraction gives $x_2' - x_1' = \gamma(x_2 - x_1)$, equivalently, $l_{O'} = \gamma l_O$. We see that the length of the rod as measured relative to the stationary frame is smaller than the length of the rod as measured relative to the moving frame wherein the rod is at rest. This phenomenon is referred to as length contraction (also Lorentz contraction, Lorentz-Fitzgerald contraction, length contraction, and simply contraction). Keep in mind that all measurements carried out by the observer in O are carried out using her lattice of clocks.

Time Dilation



The clock in O' registers t_1' when $x' = x_1$. It registers t_2' when $x' = x_2$. The clocks in O register t_1 and t_2 at the points of coincidence.

An analogous calculation for the time coordinate must yield a similar result, as space and time have close to equal footing. Here, we shall single out one of the many clocks distributed throughout the O' frame. This clock is located at x', as indicated in the sketch on the right. It ticks away at its x' position. The clock we have chosen in the O' frame passes clocks in the O frame at positions x_1 and x_2. Times are recorded in both frames at the points of spatial coincidence, namely, t_1' and t_1 when $x' = x_1$, and t_2' and t_2 when $x' = x_2$. It is then a simple matter to work out the algebra. From the Lorentz transformation, we have

$$t_1 = \gamma t_1' + \gamma (v / c^2) x' \tag{1.15}$$

$$t_2 = \gamma t_2' + \gamma (v / c^2) x' . \tag{1.16}$$

These expressions yield $t_2 - t_1 = \gamma(t_2' - t_1')$. Thus, the time interval measured in the stationary frame is larger than the time interval measured in the moving frame according to $\Delta t = \gamma \Delta t'$. This is not surprising, given that we found earlier that $t = \gamma t'$ when tracking the O' origin. The perception of the person in the stationary frame is that time passes more slowly in the moving frame. Of course, the system is symmetric, so an observer in the O' frame thinks that O' is stationary and O is in motion.

A few examples are now presented that illustrate aspects of length contraction, time dilation, simultaneity, and means of observation: lattice of clocks versus visual observation. Following this, we shall examine the "squared interval" that arises from the transformation expressed with eqns (1.2) and (1.3), namely, $\pm(c^2t^2 - x^2 - y^2 - z^2)$. Like its Euclidean 3D counterpart: $r^2 = x^2 + y^2 + z^2$, it has the same value in different inertial reference frames. At the same time, it differs greatly from its Euclidean counterpart in appearance and in the remarkable consequences it engenders.

Example 1.1. Simultaneity

Examples that deal with the issue of simultaneity in special relativity can be found in textbooks and on the web. The present example and the next one are adapted from [15]. Referring to Fig. 1.5, clocks A and B are synchronized and placed on a platform at equal distances from an observer. For example, the clocks can be brought to a common location, synchronized there, and then carried to their final locations on the platform. This can be done slowly enough to avoid relativistic effects due to their transport. Each clock sends out a short burst of light when it reads $t = 0$. The observer at rest on the platform midway between the clocks receives the light pulses. If they arrive at the same moment, the observer judges the bursts of light at clocks A and B to be simultaneous events and the clocks to be properly synchronized.

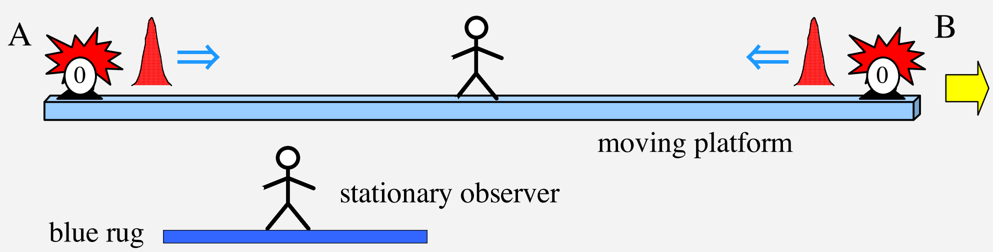


Figure 1.5. Synchronized clocks A and B emit brief pulses of light. An observer at rest midway on the platform receives the pulses simultaneously and concludes that they were emitted simultaneously. The platform is in motion relative to a second observer who judges the bursts of light as not occurring simultaneously. Adapted from [15].

Now consider the second (stationary) observer – the one standing on the blue rug. According to this observer, the clocks, the observer on the platform, and the platform are moving uniformly as a whole toward the right (yellow arrow). How will the observer on the blue rug judge the synchrony test? As far as she is concerned, the burst of light from clock A needs to traverse a greater distance to arrive at the midpoint of the platform, because the midpoint is moving away from the light pulse that originates at clock A. On the other hand, the light pulse from clock B needs to traverse a smaller distance, as the midpoint of the platform is moving towards it. Yet, the light pulses arrive at the moving midpoint at the same moment. How is this possible?

The two light pulses can only arrive at the moving midpoint at the same moment if the pulse that originates at clock A is launched earlier than would be the case for a stationary platform and a stationary observer on the blue rug. This gives the light pulse from A more time to cover the greater distance. The light pulse originating at clock B is launched later, as it needs less time to cover the distance to the midpoint.

This illustrates lack of simultaneity. The observer on the platform, being at rest with respect to the clocks, judges the flashes to be simultaneous and the clocks to be synchronized. On the other hand, the observer on the blue rug judges pulse A to happen first and clock A to be set ahead of clock B. In general, times of events must be in accord with readings of clocks that are properly synchronized, for example, by the above procedure. Judgments of simultaneity, however, in general vary from frame to frame.

Example 1.2. Apparent Rotation of Bodies in Transverse Motion

The issue of simultaneity can lead to unexpected behavior in observations that use synchronized clocks. One example is the contracted length of a moving body. Another is the body's orientation. In this example, we shall see that bodies moving in a direction that is transverse to the direction of an observer's motion appear rotated.

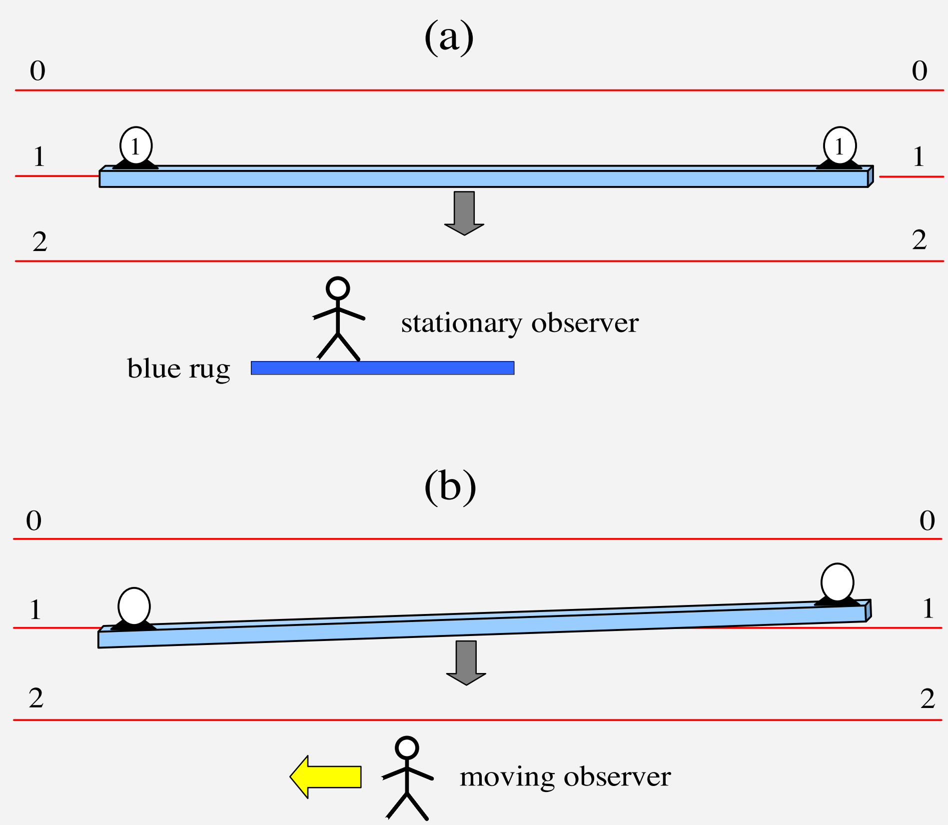


Figure 1.6. (a) The stationary observer on the blue rug sees a rod with horizontal orientation that is moving vertically downward. Numbers 0, 1, 2 indicate horizontal lines drawn in the observer's frame. (b) When the observer is moving to the left (yellow arrow), she perceives the rod as tilted with respect to the horizontal.

Figure 1.6(a) indicates a rod moving downward, passing over horizontal lines numbered 0, 1, 2, etc. The rod is parallel to the lines. Judgment of this orientation includes a judgment of simultaneity. It amounts to saying that the ends of the rod pass each horizontal line at the same moment. So when each end of the rod passes the line marked 1, the synchronized clocks at each end read the same time. Likewise when the rod passes the other horizontal lines.

Now consider an observer moving horizontally to the left, as in Fig. 1.6(b). This observer does not judge the clocks to be properly synchronized. Indeed, from Example 1.1 we see that this moving observer judges the clock on the left to be set earlier than the one on the right. In other words, the fact that the left hand side of the rod is below the line marked 1 means that it has had a longer time to travel. In other words, the left clock is set earlier. Therefore, the event of the left end of the rod passing the line marked 1, is judged by the moving observer to have happened earlier than the passing of the right end of the rod through the line marked 1. The outcome is that the horizontally moving observer regards the rod as rotated.

Again, perception of simultaneity can vary from one inertial frame to another. The tilting phenomenon demonstrated here is general in the sense that it applies to any object that moves vertically as the observer moves horizontally.

Example 1.3. Length Contraction and Apparent Rotation

Despite the facts that special relativity has been around for over a century and it is beyond reproach mathematically, paradoxical features nonetheless emerge, even for simple situations. A goal of the present example is instilling in the reader an appreciation of the care that must be exercised when interpreting different kinds of observations. For example, it drives home the fact that a lattice of clocks is the only reasonable means of judging events in gedanken experiments. Visual observation is amusing, but nuanced and not widely applicable – no match for a lattice of clocks.

Of course, lattices of clocks are fine as long as we stick to gedanken experiments. Real observations require interpretation, however. Thus, it is important to see how such observations might be interpreted.

Large speeds are needed if relativistic effects are to be manifest in the observables of classical physics. This regime is encountered is cosmology, for example, viewing the outer reaches of space through powerful telescopes. On a vastly different scale, atomic and sub-atomic phenomena must be dealt with quantum mechanically, which in chemistry translates to relativistic quantum chemistry, a common effect being that heavy nuclei bring about orbital contraction and large spin-orbit interaction. These examples are based on observation: one using a telescope to view from afar, the other using the environment – in essence the quantum mechanical version of the lattice of clocks. Here we shall examine the difference between these forms of observation.

Relativistic Cube

Consider a cube moving at high speed past a stationary observer, as indicated in Fig. 1.7. As discussed earlier, an aspect often glossed over is that of the so-called observer. Does someone visually perceive what is going on or rely on a lattice of clocks? In deriving the relativistic boost equations, the observer recorded position and time using a lattice of clocks. However, this does not in general yield the same measured quantities as those obtained through visual observation. There can be no inconsistency as long as interpretation is carried out with sufficient care.

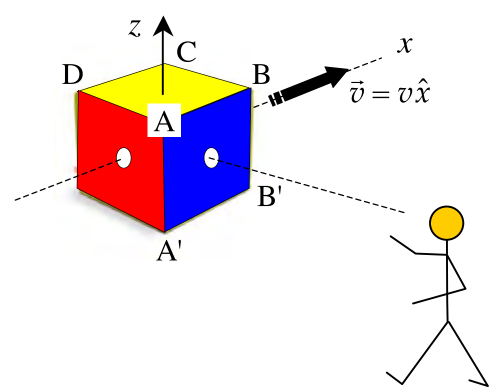


Figure 1.7. An observer puts aside her lattice of clocks. She decides instead to visually inspect the rapidly moving cube.

The fast-moving cube example serves to illustrate this point. It also serves to reinforce John Archibald Wheeler's First Moral Principle: "Never make a calculation until you know the answer." In other words, do not rush to equations before trying to figure out what is going on. We shall follow this advice and approach the present example almost entirely through discussion. Only minor algebra is needed, and even then only at the end.

Referring to Fig. 1.7, a cube that is one meter on a side, as measured in its rest frame, travels rapidly and horizontally relative to a stationary observer. For example, think of replacing the platform and its contents in Fig. 1.5 with the cube. One of the cube's surface normals is aligned with $\vec{v}$. The cube is traveling parallel to the ground and just above it, whereas the observer is standing on the ground watching the cube fly past. Figure 1.8(a) illustrates the length contraction that would be determined using a lattice of clocks. The dimension perpendicular to the page is (by definition) suppressed except for a red line that represents the red surface in Fig. 1.7.

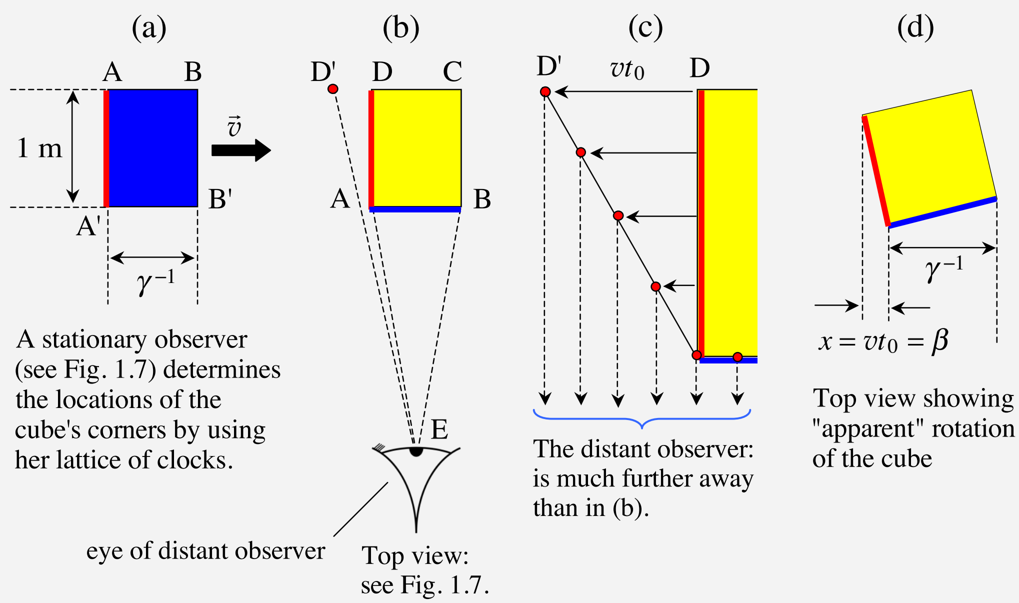


Figure 1.8. This figure adheres to the color-coding in Fig. 1.7. In (a), the red side is indicated using a red line. In (b) – (d), the top is yellow, and red and blue sides are indicated using red and blue lines. (a) Side view, as in Fig. 1.7: A stationary observer uses her lattice of clocks to determine the locations of the corners of the $1 \times 1 \times 1$ cube at a specific time, yielding the usual contraction. (b) Top view:. Photons emitted at an earlier time from the corner D of the cube, *i.e.*, when this corner was located at D', arrive at the observer's eye at the same time as those coming from the blue side. (c) Expanded view: For a distant observer, the light rays (dashed arrows) traveling from the cube to the observer are approximately parallel. The time t_0 is equal to the length of a cube side (1 m) divided by c. (d) This is what the observer perceives, namely, the red and blue sides. The distances β and γ^{-1} fit perfectly the picture of a rotated $1 \times 1 \times 1$ cube.

Referring to Fig. 1.8(b), now consider what an observer many meters from the cube would see (using one eye) at a given instant of time. This is not much different than the cosmology example of staring into deep space through a telescope. The term "instant of time" means small interval. That is, time can be measured with excellent, but finite, resolution. Because the observer is far from the cube, all light paths that originate on the blue surface and arrive at E have approximately the same length. In other words, the times required for light originating from anywhere on the blue surface to travel these distances are assumed to be the same within the time resolution of the observation.

The observer visually detects blue photons that emanate form the front side. Had she relied solely on her lattice of clocks, the shape and color indicated Fig. 1.8(a) would have been obtained.

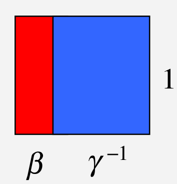
Interestingly, the observer also visually detects photons that emanate from the red side, as noted in Fig. 1.8(b) and (c). At any given instant (imagine a fast shutter isolating a short time interval δt) photons that originated at D' arrive at the eye of the observer in temporal coincidence with photons that originated at A. However, the photons that originated at D' must have been emitted earlier than those that originated at A in order for them to arrive at the observer's eye in temporal coincidence with those from A.[6] In other words, the photons coming from the upper left corner D commenced their journey to the observer at an earlier time, when D was at D'.

The observer finds that the red side of the cube is visible due to the fact that the cube is moving. Let us look further into what the observer sees visually. Figure 1.8(c) is an expanded view of the region surrounding the left side of the cube. Light rays (dashed vertical arrows) are headed toward the distant observer. Figure 1.8(c) illustrates the fact that the observer will see the entire red side of the cube, albeit contracted in an interesting way. Each point on this side emits photons at a time that enables these photons to arrive at the observer's eye in temporal coincidence with photons coming from A.

Let us now see how the numbers work out. The distance DD' in Fig. 1.8(c) is equal to vt_0, where t_0 is the "additional time" required for light that originates at D' to get to the observer. Again, the idea is that light that originates at D' should reach the observer in temporal coincidence with light that originates at A. This dictates what the observer sees visually at a given instant of time. Referring to Fig. 1.8(c), for a cube of dimension $1\times1\times1$ in its rest frame, light that departs from D' needs to travel approximately (*i.e.*, enlisting the distant observer assumption) one additional meter relative to light that departs from A. Therefore, $ct_0 = 1$, and $vt_0 = v/c = \beta$.

Figure 1.8(d) shows how these dimensions enter the perception that the cube is rotated. The observer sees two rectangles: one red, one blue. The dimensions of the rectangles are $1\times\beta$ and $1\times\gamma^{-1}$, consistent with the observer's perception that the $1\times1\times1$ cube is rotated. For example, referring to Fig. 1.8(d), the length of a side of the uncontracted cube is equal to 1. Figure 1.8(d) shows that this is equal to $\beta+\gamma^{-1}$, which is larger than 1.

Figure 1.9. Referring to Fig. 1.8(b), this is what the observer detects visually. The blue side is contracted by γ^{-1}. The red side emits photons over the projected distance $vt_0 = \beta$.

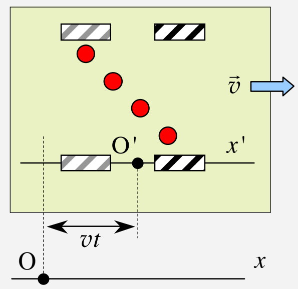


[6] In referring to photons going from A to E, it is understood that this applies to photons originating at anywhere along the line AB and going to E. All these distances are the same, to within the time resolution of the measurement, as long as the observer is sufficiently far from the cube.

View from the Cube's Rest Frame

The alternate approach presented here enables one to appreciate more fully how the shape of the $1 \times 1 \times 1$ cube enters into a "visual observation" that is carried out by a distant observer, for example, using a camera with a fast shutter. Let us return to Fig. 1.2(b), which is reproduced on the right. A stationary observer in O records a light pulse, using her lattice of clocks, that is launched in the vertical downward direction in O'. The pulse travels downward and to the right in O (red circles). The observer finds that the pulse travels a distance γl, where l is the distance between the mirrors, and $\gamma > 1$.

Fig. 1.2(b)



Suppose we had chosen instead to have the stationary observer in O observe visually a light pulse traveling vertically downward in her reference frame: a vertical line of red circles. In this case, relative to the O' frame, the pulse would have to be directed downward and toward the left. When the pulse travels a distance l in the O' rest frame of the mirror pair, the observer in O detects a pulse of contracted length $l\gamma^{-1}$.

The above scenario is now adapted to the $1 \times 1 \times 1$ cube. Referring to Fig. 1.10 (top view), photons emitted at the indicated angle in the cube's rest frame travel toward the observer at the cube's closest approach to the observer. The lengths DX = γ^{-1} and DA = 1, give XA = $(1-\gamma^{-2})^{1/2}$ = β. By inspection, AY = γ^{-1} and BY = β. This shows how the cube relates to the visual image. The observer sees a rotated cube, whether she realizes it or not. Of course with color-coding she would realize it immediately.

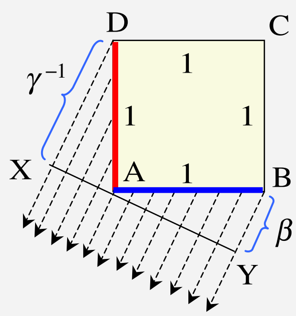


Figure 1.10. Photons that arrive at the observer (*i.e.*, comprise the XY wave front) depart at the same angle in the cube's rest frame. This shows that the observer receives photons from the bottom and side of the cube, such that the cube appears rotated.

Finally, returning to the cube's third dimension (z in Fig. 1.7), in the present example it does not undergo contraction, so it need not enter in a significant way. It would be relevant if the cube had a non-symmetrical orientation relative to the observer's frame. Even then, rather than inflicting pain on yourself by trying to work with such a situation, you could simply use the fact that the Lorentz group includes 3D rotations. In other words, rotate the cube's frame such that length contraction takes place along a convenient direction, carry out whatever analysis is called for, and then rotate back to the original orientation, if you so choose. However, we have not yet dealt with this part of the Lorentz group, so for the time being let us stick to the above graphic description.

Relativistic Sphere

The same strategy that was used in the analysis of the moving cube is now applied to a sphere that is moving with constant velocity relative to an observer. Following this exercise, you are encouraged to ponder analogous cases that can be understood through the cube and sphere examples. A number of animations of such relativistic effects can be found on the web. As with the cube, most of the time spent on the present example will be devoted to figuring out what is going on prior to carrying out any math.

When the stationary observer measures the moving sphere using her lattice of clocks, she finds that one of its dimensions is contracted. An extremely relativistic case in which the value of v is close to c is shown in Fig. 1.11. The sphere is recorded as being a rather flat oblate symmetric top in the observer's rest frame.

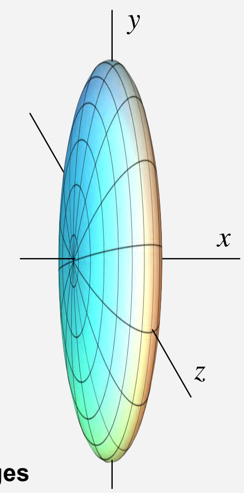


Figure 1.11. An object with a spherical shape in its rest frame moves at a speed close to c relative to an observer's rest frame. It is found to have the shape of an oblate top when the observer uses her lattice of clocks. Here the velocity is in the x-direction. A cross section, say in the xy plane, has an elliptical periphery.

Google Images

Cutting the contracted sphere in Fig. 1.11 into slices of infinitesimal thickness creates 2D images. The slices are such that each slice has one of its two symmetry axes aligned with the direction of motion, specifically, the axis for the smaller of the two ellipse dimensions. To visualize this, think of cutting the oblate top in Fig. 1.11 into slices whose flat sides lie parallel to the xy plane. Each slice displays an elliptical outline. For the analysis that follows, any one of these slices would suffice. We shall focus on the one that lies in the xy plane.

View from the Observer's Rest Frame

Figure 1.12 uses the slice of the contracted sphere that contains the origin, *i.e.*, the centermost slice – the one lying in the xy plane in Fig. 1.11. As mentioned above, the shape is elliptical. We need only obtain the solution for one slice, as the solutions for the others follow immediately. Namely, because there is no contraction in the direction of the slice normal (z direction in Fig. 1.11, into and out of the page in Fig. 1.12), the total solution is obtained by combining the solutions for the individual slices, as each of these has the same form.

It is assumed that an observer is looking up from far below, so she can only discern a line, the width of the line being the thickness of the slice. For a slice of infinitesimal thickness, this is a flatlander view.[7] The question to be answered is elementary: What is the length of the line that the observer sees?

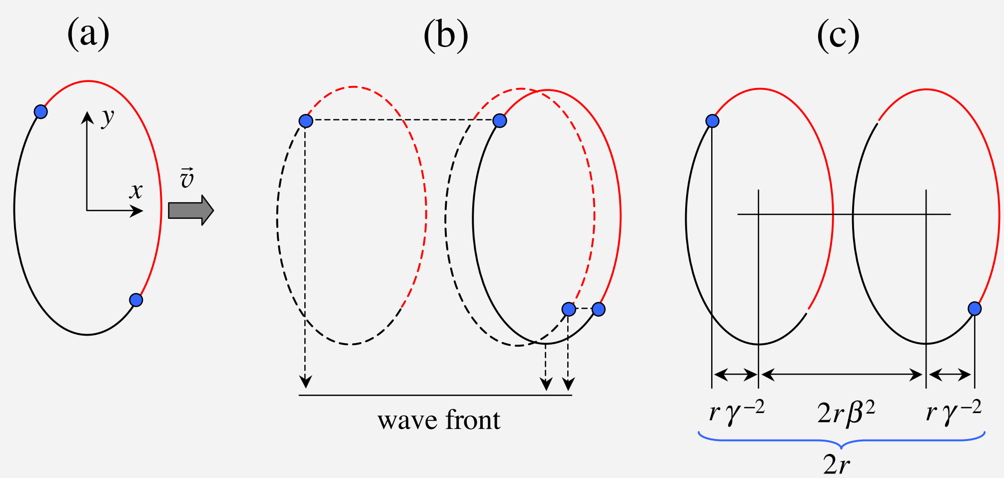


Figure 1.12. (a) In the observer's rest frame, her lattice of clocks reveals the elliptical shape of a slice of the contracted sphere. (b) Photons emitted from the uppermost point on the ellipse that permits unobstructed travel to the observer arrive at the wave front in temporal coincidence with photons that come from other points on the black portion of the periphery. These other photons are emitted at later times. The lower right blue dots indicate the other limiting case on the black part of the periphery. (c) Minor algebra shows that the length seen from below is $2r$, *i.e.*, the diameter of the uncontracted circle. The observer sees the other dimension (other slices) as uncontracted, so she visually "sees" an uncontracted but rotated sphere. She sees one side of the rotated sphere but not the other side (black versus red). For example, if the sphere had a complex surface, like the earth, this relativistic effect would enable her to view things that are hidden in the non-relativistic regime (*vide infra* Fig. 1.13).

To form a wave front that propagates toward the observer, photons that originate from different parts of the circle must travel unobstructed in a downward direction. This is obviously what happens at the bottom of the ellipse, as there is nothing to get in the way. Over a certain portion of the ellipse, however, photons cannot travel downward because they would encounter the moving ellipse.

Figure 1.12(a) illustrates, for given velocity $\vec{v}$, the region on the ellipse's periphery where photons can travel downward unobstructed. These photons originate from the black part of the periphery. On the other hand, downward trajectories of photons are blocked if the photons originate from anywhere on the red part of the periphery.

The blue points indicate demarcation points. For example, consider the blue point in the upper left of Fig. 1.12(a). The velocity $\vec{v}$ is such that photons originating to the right of this point (on the red part of the periphery) cannot move downward because the ellipse is in the way. At this blue point, $-dy/dt$ is equal to the speed of light. Note that this is a

[7] Edwin Abbott Abbott wrote the satirical novel about Victorian England *Flatland: A Romance of Many Dimensions* in 1884. It depicts life in an imaginary 2D world.

phase velocity. No material point moves at c. It is a point of intersection that moves, like when a scissors closes. Just below this blue point (*i.e.*, on the black part of the periphery), $-dy/dt$ exceeds c, so photons move downward without obstruction. The same reasoning applies to the blue point on the lower right of the ellipse in Fig. 1.12(a). Just above it (red part of the periphery), a downward traveling photon encounters the moving ellipse so it cannot reach the observer. In Fig. 1.12, large speed results in severe contraction. As $|\vec{v}|$ approaches c, the ellipse approaches a straight vertical line and the analysis simplifies. For example, the blue points appear at the top and bottom of the line.

Figure 1.12(b) illustrates this further. For photons to arrive at the observer's eye in temporal coincidence, they must arrive at the horizontal line at the bottom (wave front) at the same time. The upper horizontal dashed line indicates where emission originated from the blue point on the ellipse. The dashed ellipse at the location where this earlier emission took place reminds us that its presence there is a thing of the past. Likewise, the lower right horizontal dashed line and corresponding dashed ellipse indicate where emission took place in order that the emitted photons arrive at the wave front at the same time as the other photons that arrive at the observer's eye at her instant of observation.

Now a little algebra: Referring to Fig. 1.12, c is equated to $-dy/dt = -(dy/dx)(dx/dt) = -(dy/dx)v$. Thus: $dx/dy = -\beta$. The ellipse equation yields dx/dy. Specifically, differentiating: $r^2 = y^2 + \gamma^2 x^2$, yields $dx/dy = -\gamma^{-2} y/x$, and therefore $y = \beta\gamma^2 x$. Putting this into the ellipse equation (and making sure the correct signs are taken) yields

$$x = -r\gamma^{-2} \text{ and } y = r\beta \quad \text{(upper left demarcation point)}$$

$$x = r\gamma^{-2} \text{ and } y = -r\beta \quad \text{(lower right demarcation point).}$$

Figure 1.12(c) illustrates the different components that go into the overall length perceived by the observer. The two shorter lengths are each $r\gamma^{-2}$, while the distance between the ellipse origins is vt_0, where t_0 is the time difference for light from the upper left and lower right dots to reach the wave front. Using $t_0 = 2y/c$, this length becomes $vt_0 = 2r\beta^2$. Adding the lengths together gives $2r$, as indicated in (c). When all of the slices are added together, the outline of a full (not contracted) sphere is seen from below. However, if one were to paint things on the sphere, it would be obvious that the sphere has been rotated.

View from the Sphere's Rest Frame

As with the cube, this case is dealt with most efficiently in the sphere's rest frame. Photons that ultimately arrive at the observer must be emitted at an angle that compensates for the horizontal speed $\vec{v}$ of the sphere, as indicated in Fig. 1.13. The angle is such that the photons travel vertically in the observer's rest frame.

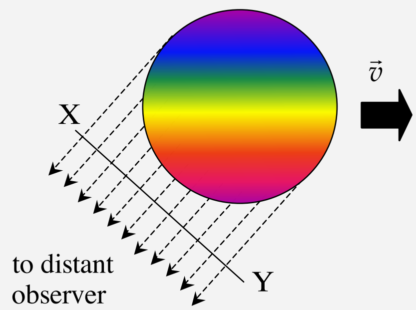


Figure 1.13. All photons that eventually arrive at the distant observer (*i.e.*, comprise the XY wave front) depart at the same angle in the sphere's rest frame. A sphere always projects to a sphere, so the visual image at the point of observation shows the outline of a sphere. However, a pattern on the surface of the sphere projects differently depending on the speed.

To summarize, the observer has two options for judging the fast moving sphere: visual inspection, but with the understanding that an instant of observation includes photons that have traveled different distances, versus a lattice of clocks, which yields a recording by using synchronized clocks distributed throughout the observer's inertial reference frame. Visual inspection is fascinating but subtle enough to require a careful look at each case. It is not easy to see the generality. Thus, we shall stick to the lattice of clocks.

Example 1.4. Time Dilation and Muon Decay

It is well known that when cosmic rays enter the outermost regions of the earth's atmosphere muons are created in abundance. These are negatively charged particles (leptons) whose mass is about 200 times that of an electron. Some of these muons travel toward the earth at a speed near c. We shall consider muons having a speed of $0.98c$. They are relatively unimpeded as they travel through the atmosphere with a speed near c. An interesting and important feature of muons is that they are unstable. That is, they decompose into fragments. In the muon rest frame, it decays to three particles: an electron, an electron antineutrino, and a muon neutrino with an exponential lifetime of approximately 2.2 μs.

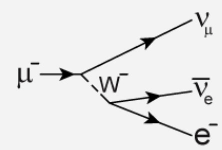


Figure 1.14. The muon is an elementary particle (no internal structure). At one time it was believed to be a meson.

An experimentalist employed by CERN (an international research facility not far from Geneva, Switzerland) was working at the top of a nearby mountain whose altitude is 3000 m. She measured 1000 muons. She then traveled to sea level where, under otherwise identical conditions, the measurement was repeated. At a speed of $0.98c$, the time required to travel 3000 m is approximately 10.2 μs. Had this value been used with the exponential lifetime of 2.2 μs to estimate the number of muons that reach sea level, the anticipated number of muons would have been $1000\exp(-10.2/2.2) \approx 10$. However, this is not at all what was recorded. Instead of 10, the number of detected muons was 380 – a remarkably larger number of muons than 10. What is going on?

This is an example of the time dilation phenomenon. In the muon's rest frame the amount of time that has elapsed during the trip from the top of the mountain to sea level is not 10.2 μs, but 10.2 μs$/\gamma$. When γ is calculated using $v = 0.98c$, the value $\gamma = 5.03$ is obtained. Thus, instead of 10.2 μs, a time of 2.04 μs must be used to estimate the number of muons that survive the trip. This yields $1000\exp(-2.03/2.2) = 400$, in good agreement with the measurement of 380 when the experimental uncertainty is taken into account.

Discussion: About Photons (Arman, Pavel)

This discussion arose from a question posed by Arman and comments from Pavel and others about the measurement of relative speed. It turned out that the issue was one of distinguishing relative speed from "mutual speed." At the time, I had not heard of mutual speed. Subsequently I read about it and found that it is amusing but not a big deal. It is discussed in Section 3.5. This synopsis captures the gist and spirit of the discussion.

Photons do not have velocity in the same sense as massive particles. In relativistic classical mechanics, particle velocity is defined as $U^\alpha = dx^\alpha / d\tau$, where x^α is a space-time component, with α denoting its coordinate, and τ is the proper time, the time measured in the particle's rest frame. A photon has no rest frame, so such a definition is meaningless for a photon. It makes no more sense than trying to catch a photon by running after it ever faster.

Packets of photons can be prepared that are tightly bunched in space and time. They propagate in vacuum with speed c without spreading (leaving aside diffraction). This lack of spreading is due to the fact that the constituent waves, whose momenta span a broad range, each have phase velocity c. Despite their highly localized nature, these packets are photons: no rest frame, the same speed in all frames, and so on. A single photon can be prepared in such a packet.

At the other extreme, a photon can have well defined momentum. As the distribution of its k values narrows, its momentum, for all practical purposes, becomes known precisely. In this limit, a photon can be expressed in four-vector form (using $\hbar = 1$) as

$$\begin{pmatrix} \omega / c \\ k_x \\ k_y \\ k_z \end{pmatrix}. \tag{1}$$

Frequency and wave vector satisfy $\omega^2 = c^2 k^2$. This follows from elementary electromagnetism, and it is consistent with eqn (1). We have not yet seen how this follows mathematically from eqn (1), but this is coming soon.[8] As with any Lorentz covariant four-vector, Lorentz transformation retains the four-vector's form, but changes the perspective insofar as components to that of a different reference frame. Applying Lorentz transformation to eqn (1) with $k_y = k_z = 0$ yields

$$\begin{pmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} \omega / c \\ k_x \end{pmatrix} = \begin{pmatrix} \omega' / c \\ k_x' \end{pmatrix} = \begin{pmatrix} \gamma\omega / c + \beta\gamma k_x \\ \beta\gamma\omega / c + \gamma k_x \end{pmatrix}. \tag{2}$$

The column vector on the far right gives both the frequency shift due to the Doppler effect and the new wave vector:

---

[8] A Lorentz scalar is obtained by contracting the energy-momentum four-vector, yielding $P^\alpha P_\alpha = (\omega / c)^2 - k_x{}^2 - k_y{}^2 - k_z{}^2 = (mc)^2$. This vanishes for $m = 0$, leaving $\omega^2 = c^2 k^2$.

$$\omega' = \omega\gamma(1+\beta c k_x / \omega) = \omega\gamma(1+\beta) \tag{3}$$

$$k_x' = k_x\gamma(1+\beta\omega / c k_x) = k_x\gamma(1+\beta)\,. \tag{4}$$

Note the constancy of the speed of light c as frequency and wave vector change synchronously under Lorentz transformation:

$$\frac{(\omega')^2}{(k_x')^2} = \frac{\omega^2}{k_x{}^2} = c^2\,. \tag{5}$$

When a photon's $\vec{k}$ value is sufficiently well defined, the photon spans a large spatial region. Distances of tens of meters at infrared wavelengths were easily achieved when I was in graduate school. There is nothing to prevent photons that have the same $\vec{k}$ value (and therefore the same value of ω) from being in the same place at the same time, as photons are bosons. In fact they prefer this. This is exactly what happens with lasers, where many photons (let us say 10^{20}) occupy a single cavity mode, with their coherence (oscillation phase with respect to one another) assured through the stimulated emission process. The presence of photons in a specific cavity mode motivates photons that are being created to enter this mode. In effect, the photons combine to form a classical electromagnetic wave.

Now consider an experiment in which tightly bunched photon packets move in the same direction on a common path (leaving aside diffraction). Each packet moves at the speed of light, with the distances between them remaining fixed. Though the packets cannot catch one another, if we adjust the time delay between two of them, these two packets can be brought into spatial and temporal coincidence. If their phases are adjusted at the same time, the resulting single pulse can retain the shape of the original pulses. The photons do not catch one another. Rather, the phases of their constituent modes are arranged such that the temporal shape of the tightly bunched packet is preserved. Said differently, the photons occupy the same free space modes.

Suppose that at $t = 0$ an observer sends out two ultrafast packets of photons (let us say of 10 fs duration) traveling in opposite directions. The observer finds that the distance between these packets grows at a speed of $2c$. However, this is a pair of measurements. The same idea applies to massive particles. One particle can move away from an observer at a speed close to c, while another particle can move away from the observer in the opposite direction, also at a speed close to c. The distance between them grows at a speed of almost $2c$. However, this is entirely compatible with c being the universal speed limit.

Were an observer moving very rapidly (0.9999 c) along the same path as a photon packet, no matter how fast she traveled, she would always find that the pulses travel at the speed of light with respect to her. No matter how small the mass, or how high the particle energy (imagine a highly relativistic electron neutrino, whose mass is believed to be a mere ~ 1 eV), the massless photons travel at the speed of light with respect to the massive particle. This is the nature of massless versus massive particles in inertial reference frames. This changes when the observer is undergoing acceleration. For example, photons might never catch up to an accelerated observer from the perspective of her rest frame. Amazing!

2. Invariant Interval

The coordinates ct, x, y, and z can be placed in a 4D column vector that is analogous to the 3D one used for the Cartesian components x, y, and z. This 4D vector is written

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}. \tag{2.1}$$

In the 3D Cartesian case, the quantity $r^2 = x^2 + y^2 + z^2$ is invariant with respect to a change of basis. It is a scalar: the squared length of a vector that originates at the origin and terminates at the point (x, y, z). Like the vector itself, r^2 is unaffected by the choice of coordinate system that is used to represent the vector. On the other hand, in the case of 4D spacetime, if a scalar is calculated from eqn (2.1) by multiplying row times column, the quantity that is obtained: $(ct)^2 + x^2 + y^2 + z^2$, is not invariant with respect to a change of reference frame brought about by a Lorentz boost. Rather, the Lorentz invariant quantity is $S^2 = (ct)^2 - x^2 - y^2 - z^2$.

The *ict* Coordinate of Poincaré

The Lorentz transformation can be applied to a vector that consists of three spatial coordinates plus one that is proportional to ct. Following Einstein's 1905 paper, his former teacher, Hermann Minkowski, argued in 1908 in favor of a four-vector that includes time. Einstein did not initially accept this idea, but eventually came round. Henri Poincaré already in 1905 had suggested that ict be assigned this role (x, y, z, ict), as this would enable $-S^2$ (also a conserved quantity) to be obtained from a dot product: $x^2 + y^2 + z^2 + (ict)^2$. This idea caught on, and for many years ict was used. However, by ~ 1980 it had been disparaged, mainly because of shortcomings in general relativity. You will only find ict in older texts. The modern approach is to transfer the role of i to a basis vector through the use of a metric tensor. This enables the math to be carried out in a way that is compact, unambiguous, and in keeping with the theory of general relativity and other advanced theories. See Kusse and Westwig [13] for an accessible and reasonably thorough discussion.

To see how the invariance of $S^2 = (ct)^2 - x^2 - y^2 - z^2$ works, transform it to the primed system by using eqn (1.2) and the fact that $y' = y$ and $z' = z$.

$$S^2 = (ct)^2 - x^2 - y^2 - z^2 \tag{2.2}$$

$$= \gamma^2 \left(ct' + \beta x'\right)^2 - \gamma^2 \left(x' + \beta ct'\right)^2 - (y')^2 - (z')^2 . \tag{2.3}$$

$$= \gamma^2 \left(\left(1-\beta^2\right)\left(ct'\right)^2 - \left(1-\beta^2\right)\left(x'\right)^2 \right) - (y')^2 - (z')^2 \tag{2.4}$$

$$= (ct')^2 - (x')^2 - (y')^2 - (z')^2 \,. \tag{2.5}$$

This demonstrates that S^2 is invariant with respect to the Lorentz boost. It is also invariant with respect to 3D rotation, which is obvious because rotation leaves $x^2 + y^2 + z^2$ unchanged. Thus, S^2 is invariant with respect to the Lorentz transformation of interest in this chapter: boost plus rotation.

The quantity S^2 is referred to as the squared interval, or simply the interval.[9] It is defined here as: $S^2 = (ct)^2 - x^2 - y^2 - z^2$, but sometimes it is defined as the negative of this: $S^2 = -(ct)^2 + x^2 + y^2 + z^2$. Clearly, each is conserved in a Lorentz transformation. Different scientific communities prefer one or the other convention, with 70% preferring the former. They are equivalent, but going between journal articles and texts that use different conventions is annoying. The fact that S^2 can have a value that is positive, negative, or zero justifies use of the term interval. When S^2 represents a wave front that advances at the speed of light, $S^2 = 0$, so $(ct)^2 = x^2 + y^2 + z^2$. No material object or point can travel faster than c, so the distance it travels, $(x^2 + y^2 + z^2)^{1/2}$, is less than ct. Here the capital letter S indicates the interval from (0,0,0,0) to (ct, x, y, z). Later, S will be replaced with small s to denote pairs of points: $(ds)^2 = c^2(t_2 - t_1)^2 - (x_2 - x_1)^2 - (y_2 - y_1)^2 - (z_2 - z_1)^2$, as pointed out in the footnote, including infinitesimals: $(dct)^2 - (dx)^2 - (dy)^2 - (dz)^2$.

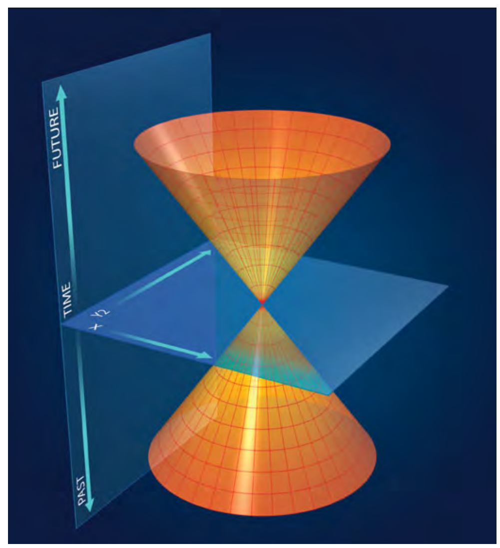


Figure 2.1. The light cone is obtained from 4D spacetime by suppressing one of the spatial coordinates. The cones at 45° correspond to the speed of light. With the observer at the origin (apex), her past lies inside the lower cone, whereas her future lies inside the upper cone. (Google Images.)

A surface defined by the equation: $(ct)^2 - x^2 - y^2 - z^2 = 0$, would require 4D visualization – a skill that eludes our species. However, by suppressing one of the spatial coordinates (set it to zero), the light cone in Fig. 2.1 is constructed. Dynamical processes are represented within the upper and lower light cones. This means that events can be related causally. That is, the time elapsed between them

[9] The origins O and O' are reference points that enable a coordinate to be treated as a vector. For example, in 3D the points (x, y, z) can define a vector that goes between the origin $(0, 0, 0)$ and (x, y, z). In this sense, the point is used to define an interval. Later, the term spacetime interval will be used in a more general sense, namely, to describe the difference between two events: (ct_1, x_1, y_1, z_1) and (ct_2, x_2, y_2, z_2).

exceeds the distance between them divided by c. For example, for a particle at the apex of a double cone, the particle's past, present, and future dynamics lie within the cones.

As a particle wends its way through spacetime on its world line, it can be thought of as carrying a light cone around with it. No matter how the particle's world line twists and turns, there is always a vertical light cone, with the particle at the apex. If the value of $(ct)^2 - x^2 - y^2 - z^2$ [assuming the apex is (0,0,0,0)] in the immediate vicinity of the apex is positive, the curve is said to be timelike in this region of spacetime, and the events it describes are related causally. All dynamical processes are causally related. The three light cones on the world line of course are drawn in hindsight. As the particle makes its way, its record of events are for the past and present, obviously not the future.

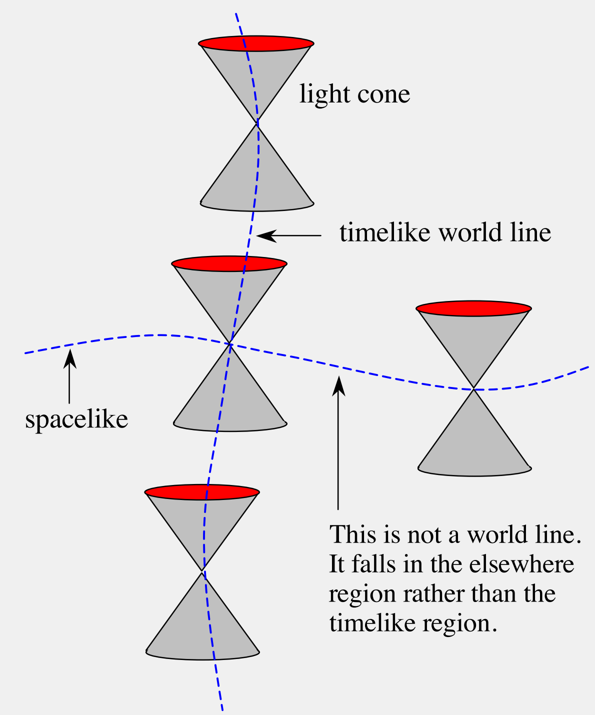


Figure 2.2. A particle's world line has slope < 45° with respect to the vertical. The spacelike path can be that of a point, but not of a particle or photon. A spacelike ray is measured with a lattice of clocks. The only physical point that can be spacelike belongs to phase velocity.

If $(ct)^2 - x^2 - y^2 - z^2$ is zero, we have the speed of light. If it is negative, events are not causally related. For example, a lattice of clocks that records simultaneous events is an extreme case in that $\Delta t = 0$. It is possible that a (massless) point follows a spacelike world line. This applies to phase velocity, a simple example being the intersection point on a pair of scissors. Phase velocity is discussed in Appendix 1.

Reducing the dimension from 4D to 3D or 2D facilitates visualization and also has a sound basis. For example, particles have well-behaved world lines. Of course, they experience forces, so their world lines are not straight. In the immediate vicinity of an apex, however, the spatial direction along which motion takes place can be assigned a Cartesian coordinate, yielding a 2D representation. The 3D diagram has more visual appeal, but nothing is lost with a 2D slice. Think of a vertical sheet through Fig. 2.2 that includes the origin.

2.1. Passive Transformations: Minkowski Space

There are two equivalent ways to effect a boost transformation between stationary and moving (unprimed and primed) frames. These are referred to as passive and active. In the passive version, a point – say (ct, x) in the unprimed frame – is held fixed, and transformation of the unprimed frame creates a new (primed) reference frame. The point (ct, x) has coordinates (ct', x') relative to the primed frame.

Rotation of a 3D Cartesian frame is easy to visualize, as axes rotate in the same sense. Think of 2D rotation of x- and y-axes by some angle; nothing could be simpler. On the other hand, axes rotate in opposite senses with a Lorentz boost, as indicated in Fig 2.3. Whereas sines and cosines appear when rotating the axes in the same sense, hyperbolic functions appear in the boost case.

The history surrounding the hyperbolic geometry of the Lorentz group (boosts and rotations) is rich with anecdote, personality, and intrigue, reflecting the excitement of the time. In this section, we shall touch on a limited number of fundamental aspects and considerations. I happened upon an excellent article on this topic, including some interesting history [24]: J. A. Rhodes and M. D. Semon, *Relativistic Velocity Space, Wigner Rotation, and Thomas Precession*, Am. J. Phys. **72**, 943-960 (2004). It is nicely written and accessible to non-specialists.

Figure 2.3 illustrates the (ct,x) plane of a (stationary) O frame and the non-orthogonal ct' and x' axes of a (moving) O' frame. As before, it is assumed that the frames coincide at $t = t' = 0$. The blue line at 45° corresponds to light trajectories; it is called the light line. This figure (minus the primed axes) is the upper right quadrant of a vertical slice through a light cone that contains the cone's apex.

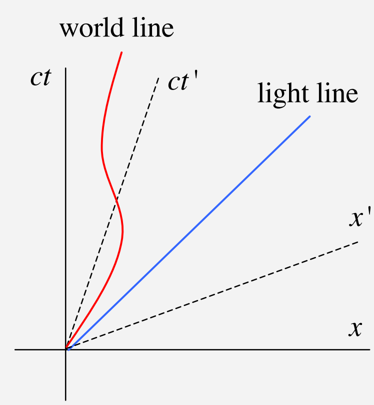


Figure 2.3. O' is in motion with respect to O. At $t = t' = 0$ the O and O' origins coincide, as indicated. The world line of a particle moving in the unprimed frame is shown in red. The world line of the O' origin advances in the unprimed frame along the ct' axis.

Referring to Fig. 2.3, a timelike path in the region of spacetime between the ct axis and the blue line is called a world line. The world line need not pass through the origin as long as its slope dx/dt is less than c. Any timelike curve in the (ct,x) plane can be interpreted as motion projected onto the x-axis. In the (ct,x) plane, the world line of the origin O' (think of the origin O' as a particle moving in the (ct,x) plane) coincides with the ct' axis, because, by definition, the O' origin remains at $x' = 0$. The constancy of the speed of light in all inertial frames requires that $\Delta x/\Delta ct = \Delta x'/\Delta ct' = 1$ for a light ray.

A point with constant velocity in the stationary frame O has a straight-line path in a ct versus x diagram. As mentioned earlier, regardless of where the constant velocity points in space, it is always possible to introduce a Cartesian coordinate system such that the direction of motion coincides with its x-axis. For example, this is easily achieved through rotation of the original reference frame. To determine the direction of the x' axis, recall the constancy of c in all reference frames and the fact that $S^2 = 0$ for the light line in each frame. This requires that x' and ct' are oriented symmetrically about the light line, as indicated in Fig. 2.3.

2.1.1. Scaling the Axes

Nonzero values of the Lorentz invariant quantity $S^2 = (ct)^2 - x^2$ can be used to illustrate the fact that the (ct', x') axes do not have the same scales as the (ct, x) axes in diagrams such as the one in Fig. 2.3. Specifically, measurements that yield identical distances along the x and x' directions in the O and O' frames, respectively, corresponds to different lengths as measured by setting a ruler along the x and x' axes in Fig. 2.3. For example, a length of one meter measured in the O frame might correspond to 1 cm on the x-axis, whereas a length of one meter measured in the O' frame might correspond to a distance of 1.1 cm on the x'-axis. This scaling of axes is discussed in detail in Kusse and Westwig [16]. In addition, S^2 also serves to illustrate that measurements of the same event, when carried out in frames O and O', can yield seemingly different results due to the demise of simultaneity when viewing events from different inertial frames.

Figure 2.4 shows plots of the four hyperbolas: $S^2 = \pm 1 = (ct)^2 - x^2$. In making the figure, I cheated by using parabolas instead of hyperbolas. Proper hyperbolas can be added later. In general a timelike curve in the (ct, x) plane can be interpreted as representing the x-component of a point's motion. A particle's speed is always less than c, which requires slopes $dx/dt < c$. Thus, the upper and lower hyperbolas are not allowed as particle world lines because dx/dt exceeds c. At least as egregious a flaw insofar as these hyperbolas being particle world lines is the fact that, in going from large negative values of x to large positive values of x, time both increases and decreases. The latter is preposterous.

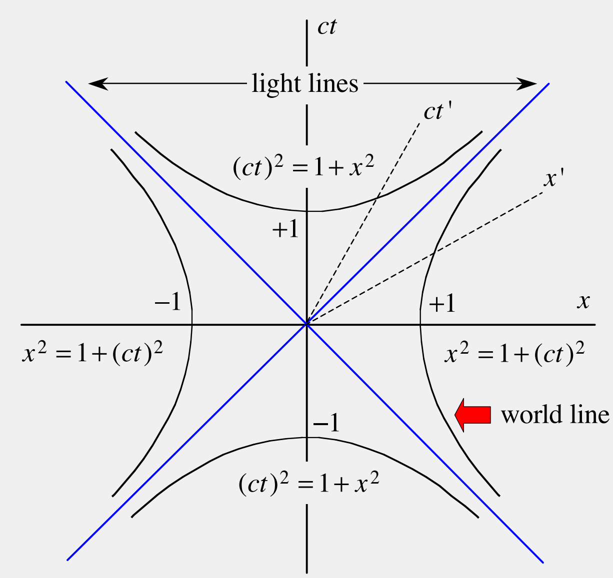


Figure 2.4. Calibrating hyperbolas are given by $S^2 = (ct)^2 - x^2 = \pm 1$. Ones labeled $x^2 = 1 + (ct)^2$ can represent a particle's world line, as velocity is $< c$.

Let us now use S^2 to elucidate the different scales of the x and x' axes. The choice $S^2 = (ct)^2 - x^2 = 1$ (the use of 1 is for convenience), gives $ct = 1$ when $x = 0$. This curve is labeled $(ct)^2 = 1 + x^2$ in Fig. 2.4. This squared interval retains its value of 1 in any primed-axes system. The number of such skewed axes is infinite, each corresponding to a different value of v.

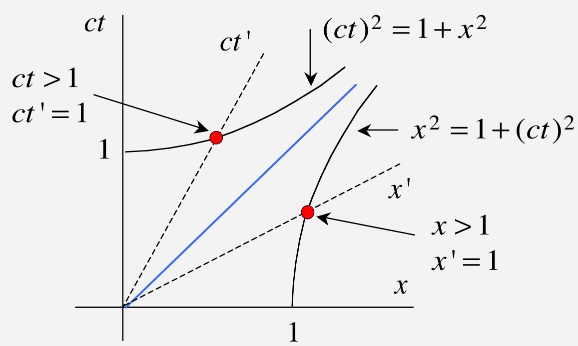


Figure 2.5. O' is in motion with respect to O. The world line of the point O' moving in the O frame ($x = vt$) coincides with the ct' axis. The 1's were chosen for convenience.

Referring to Fig. 2.5, the hyperbola $(ct)^2 = 1 + x^2$ intersects the ct' axis at $x' = 0$. This intersection point is indicated by a red dot. The invariance of S^2 means that $(ct')^2 = 1 + (x')^2$, in which case $ct' = 1$ at $x' = 0$. This tells us that the ct and ct' axes do not have the same scale. That is, a line that starts at the origin and goes along the ct axis to the point labeled 1 has a length that is shorter than the length of a line that starts at the origin and goes along the dashed ct' axis to the red dot.

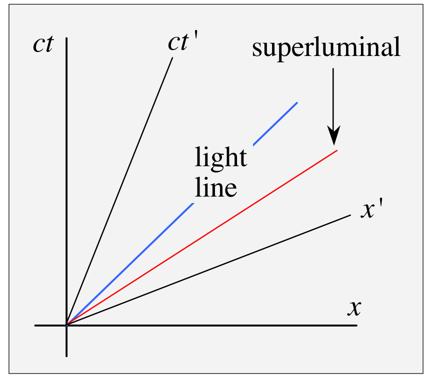
Now consider the scaling of the x and x' axes. This time we shall use the hyperbola: $(ct)^2 - x^2 = -1$. As before, the choice of -1 is for convenience. The result does not depend on whether we use -1 or any other negative number. Referring to the $x^2 = 1 + (ct)^2$ hyperbola, we see that $x = 1$ at $ct = 0$. Following the hyperbola from $x = 1$ to the x' axis (red point), we see that $x' = 1$ at $ct' = 0$, whereas $x > 1$ at this red point. Thus, the x' and x axes do not have the same scale. Note that this result could have been obtained by using the fact that the axes need to be symmetric about the light line to ensure the constancy of c in any inertial frame.

2.1.2. Time Dilation and Length Contraction Again

Expressions for time dilation and length contraction were derived in Section 1.5. It is noted that these phenomena also manifest in Fig. 2.5. The upper left red dot shows that an observer in the moving frame records $ct' = 1$ at $x' = 0$ for a given event, whereas an observer in the stationary frame records $ct > 1$ and $x > 0$ for this event. Thus, $(ct)^2 = 1 + x^2$, meaning that $t > t'$. From the perspective of the stationary observer, time passes more slowly in the moving frame than in the stationary one. Using $x = vt$, the above equation is $(t')^2 = t^2 - \beta^2 t^2$. Again we have $t = \gamma t'$, the expression for time dilation.

Now imagine a rod that is stationary in the unprimed frame and extends from $x = 0$ to $x = 1$. The world line of its tip at $x = 1$ is a vertical line. This line crosses the x' axis at a location $x' < 1$. This is an example of length contraction.

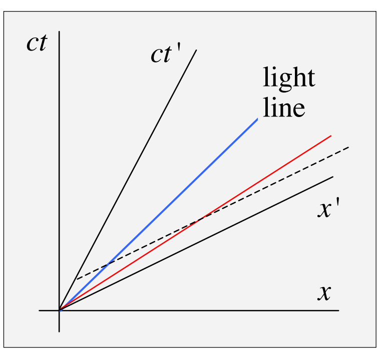
Example 2.1. Superluminal Motion



A point in space, as distinct from a particle or photon, can move at speeds exceeding c. This falls under the heading of phase velocity. Appendix 1 discusses phase velocity as well as group velocity and dispersion. The speed associated with a phase velocity can exceed c by a great deal, hence the descriptive term "superluminal motion." Standard examples are the intersection point of a pair of scissors, waves incident on a shore at near-normal incidence, and the intersection point of a beam from a laser pointer as it is swept rapidly across a wall. Such cases do not involve the transport of mass, and in the case of photons no information is transmitted at a speed exceeding c. The sketch on the right indicates a constant superluminal speed (red line) relative to the stationary (unprimed) axes.

It is interesting, if not downright amusing, to see how the above superluminal motion is perceived from the perspective of an inertial frame that is in motion relative to the stationary frame. Referring to the above sketch, this moving frame has skewed ct' and x' axes. Let us see what happens when the velocity of the moving frame is increased, causing the primed axes to collapse toward the light line.

On the one hand, you might expect the superluminal speed relative to the primed frame to *decrease* as v increases. After all, the primed frame and the superluminal point move in the same direction relative to the unprimed system. Thus, it seems reasonable to assume that the superluminal point slows relative to the primed frame as v increases.



On the other hand, the sketch on the left indicates that this does not happen. In fact, quite the opposite occurs. As x' approaches the red superluminal line, the speed of a superluminal point relative to the primed axes actually increases. The distance from the origin to a point on the red line, when projected onto the x' axis, changes little as the x' axis approaches the red line. In marked contrast, however, the projection of this distance onto the ct' axis gets smaller and in fact approaches zero as x' converges to the red line. Thus, the superluminal speed approaches infinity when observed from the primed frame. When the x' axis moves past the red line, the superluminal motion still has a large magnitude, but it now moves in the opposite direction.

Exercise: (a) Explain why there is no transfer of information when the intersection point of a beam from a laser pointer is swept rapidly across a wall. (b) Why does the superluminal motion diverge when the x' axis and red line are coincident? Explain why the superluminal motion changes direction. Think about simultaneity.

2.2. Relating Velocities Between Frames

To see how a particle's velocity is perceived from stationary versus moving frames, consider Fig. 2.6. It is assumed that the primed frame moves in the $+x$ direction with speed v ($\vec{v} = v\hat{x}$) relative to the unprimed frame. A point (red dot) is moving (also in the $+x$ direction) with speed $\mathsf{v} = \Delta x / \Delta t$ relative to the stationary frame. We wish to determine the point's speed relative to the primed frame. The coordinate displacements Δx and Δct are indicated with thin black arrows on the right side of Fig. 2.6.

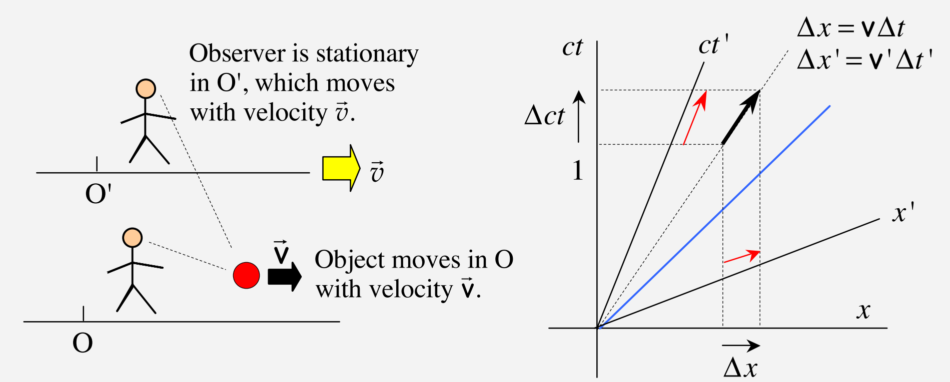


Figure 2.6. The velocity of an object relative to the stationary (lower) frame differs from the velocity of the object relative to the moving (upper) frame.

In the moving frame, the point's velocity is determined through its parallel-axis projections onto the x' and ct' axes: $\mathsf{v}' = \Delta x' / \Delta t'$ (thin red arrows). Keep in mind that the x' and ct' axes are scaled differently than the x and ct axes. Consequently, Fig. 2.6 is helpful, but not quantitative because of this last point. It is accurate, however, insofar as the ratio of the lengths of the red arrows.

To see how the speeds v' and v are related to one another, use $\mathsf{v}' = \Delta x' / \Delta t'$ and $\mathsf{v} = \Delta x / \Delta t$ with the Lorentz transformation given by eqn (1.3): $\Delta x' = -\beta\gamma\Delta ct + \gamma\Delta x$ and $\Delta ct' = \gamma\Delta ct - \beta\gamma\Delta x$, to write

$$\mathsf{v}' = \frac{\Delta x'}{\Delta t'} = c\,\frac{\Delta x - \beta\Delta ct}{\Delta ct - \beta\Delta x} \tag{2.6}$$

Dividing the numerator and denominator by Δt yields

$$\mathsf{v}' = c\,\frac{\mathsf{v} - \beta c}{c - \beta\mathsf{v}} \tag{2.7}$$

$$= \frac{\mathsf{v} - v}{1 - v\mathsf{v}/c^2}\,. \tag{2.8}$$

We see that when the speed of the primed frame relative to the stationary frame is equal to the speed v of the point in the stationary frame, the primed speed v' vanishes, as it must. When v exceeds greatly the speed of the moving frame, v, and it approaches c, v' approaches c, again as it must. In this section, velocities were parallel, simplifying the math. The more general case will be considered later.

2.3. Matrix Representation: Velocity Addition

The matrix representation for the Lorentz boost given by eqn (1.3), and using $y' = y$ and $z' = z$, is

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}. \tag{2.9}$$

Such representation is useful when dealing with relative motion that involves more than two inertial frames, including boosts in different directions. An example of relative motion was seen in the last section. Here we shall derive an analogous result.

The compounding of relative motions for more than two reference frames can be obtained through successive application of Lorentz boosts using matrix multiplication. For example, suppose a double-primed frame O'' moves with constant velocity v' with respect to the single-primed frame O', which in turn moves with constant velocity v with respect to the stationary frame O. Furthermore, assume that motion is along a common axis. This last assumption is a great simplification. The situation is more interesting, though algebraically tedious, when the successive boosts correspond to different directions.

We shall now carry out the transformation from O'' to O. This is trivial in the nonrelativistic limit. The speed is simply $v + v'$. In the case of relativity, the transformations are sequential: O'' $\rightarrow$ O' followed by O' $\rightarrow$ O, so the result is obtained by matrix multiplication using Lorentz boost matrices (labeled using capital lambda, Λ) for the relative motion of reference frames along a common axis. This time let us start with the z-axis, just to be different for a change. In labeling the matrices, the operation O'' $\rightarrow$ O' is assigned the subscript 1 because it happens first, while O' $\rightarrow$ O is assigned the subscript 2. Thus, the overall process is

$$\Lambda_2 \Lambda_1 = \begin{pmatrix} \gamma_2 & 0 & 0 & -\beta_2\gamma_2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\beta_2\gamma_2 & 0 & 0 & \gamma_2 \end{pmatrix} \begin{pmatrix} \gamma_1 & 0 & 0 & -\beta_1\gamma_1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\beta_1\gamma_1 & 0 & 0 & \gamma_1 \end{pmatrix} \tag{2.10}$$

$$= \begin{pmatrix} \gamma_1\gamma_2(1+\beta_1\beta_2) & 0 & 0 & -\gamma_1\gamma_2(\beta_1+\beta_2) \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\gamma_1\gamma_2(\beta_1+\beta_2) & 0 & 0 & \gamma_1\gamma_2(1+\beta_1\beta_2) \end{pmatrix} \tag{2.12}$$

$$= \begin{pmatrix} \gamma & 0 & 0 & -\beta\gamma \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\beta\gamma & 0 & 0 & \gamma \end{pmatrix}. \tag{2.13}$$

The parameters γ and β in the last matrix are for the overall process. By comparing matrix elements we see that

$$\gamma = \gamma_1\gamma_2(1+\beta_1\beta_2) \tag{2.14}$$

and

$$\beta\gamma = \gamma_1\gamma_2(\beta_1+\beta_2). \tag{2.15}$$

The parameter γ_1 is in terms of the velocity v' of O'' relative to O':

$$\gamma_1 = \frac{1}{\sqrt{1-(v'/c)^2}} = \frac{1}{\sqrt{1-\beta_1{}^2}}. \tag{2.16}$$

Likewise, the parameter γ_2 is in terms of the velocity v of O' relative to O:

$$\gamma_2 = \frac{1}{\sqrt{1-(v/c)^2}} = \frac{1}{\sqrt{1-\beta_2{}^2}} \tag{2.17}$$

Clearly, γ compounds the two transformations such that the velocity of O'' relative to O (labeled v'') can be obtained. Let us carry out this algebraic exercise step-by-step, starting with $\gamma = \gamma_1\gamma_2(1+\beta_1\beta_2)$:

$$\gamma = \frac{1+\beta_1\beta_2}{\sqrt{1-\beta_1{}^2}\sqrt{1-\beta_2{}^2}} \tag{2.18}$$

$$= \frac{1+\beta_1\beta_2}{\sqrt{1+\beta_1{}^2\beta_2{}^2-\beta_1{}^2-\beta_2{}^2}} \tag{2.19}$$

$$= \frac{1}{\sqrt{\dfrac{1+\beta_1{}^2\beta_2{}^2+2\beta_1\beta_2-\beta_1{}^2-\beta_2{}^2-2\beta_1\beta_2}{(1+\beta_2\beta_1)^2}}} \tag{2.20}$$

$$= \frac{1}{\sqrt{1-\left(\dfrac{\beta_1+\beta_2}{1+\beta_1\beta_2}\right)^2}} \tag{2.21}$$

$$= \frac{1}{\sqrt{1-\left(\dfrac{1}{c}\dfrac{v+v'}{1+vv'/c^2}\right)^2}} \tag{2.22}$$

$$= \frac{1}{\sqrt{1-(v''/c)^2}} \ . \tag{2.23}$$

Thus, the formula for adding parallel velocities in the relativistic regime has been obtained. The velocity of O'' relative to O is

$$\boxed{v'' = \frac{v+v'}{1+vv'/c^2}\,.} \tag{2.24}$$

Notice that this is similar to eqn (2.8) in Section 2.2.

Example 2.2. The Moving Water Experiment of Fizeau

A French scientist named Hippolyte Fizeau carried out an interesting experiment way back in 1851 that turned out to have a great deal of significance half a century later in the context of special relativity. In fact, it proved central to the development of the theory. Like the Michelson-Morley experiment, an unanticipated result demanded reconciliation. Einstein regarded the Fizeau experiment as seminal, and one of the principle motivating factors in his development of the theory. He claimed to know relatively little about the Michelson-Morley experiments.

The Fizeau experiment was the epitome of simplicity. Water was flowed through tubes, and rays of light passed through the flowing water, as indicated in Fig. 1. It was believed at the time that light passed through a "luminiferous ether" that was present everywhere. This is the ether theory you no doubt have heard about. This theory predicted that the speed of light relative to the laboratory, w, for light passing through the flowing water is given by $w = c/n \pm v$, where n is the index of refraction, c/n is the speed of light in stationary water, and v is the velocity of the flowing water. The ± indicates the direction of flow relative to the wave vector of the light: aligned with it versus opposing it. Fizeau took into account the directions of the red and blue arrows relative to the direction of water flow. He found that w obeyed the equation:

$$w = c/n \pm v\left(1 - n^{-2}\right), \tag{1}$$

which differs significantly from $c/n \pm v$.

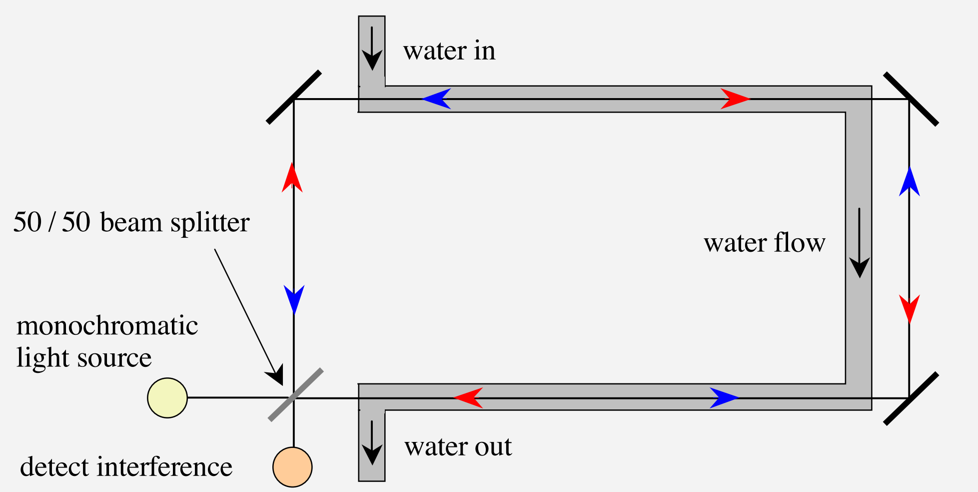


Figure 1. A 50/50 beam splitter enables monochromatic light rays to pass through flowing water in opposite directions (red and blue arrows). The beams recombine on a detector that registers their interference.

This result was a source of confusion for many years. Preposterous assumptions were needed to reconcile it within the framework of the ether theory. It was explained immediately, however, within the framework of special relativity. To see how this works, we use the fact that c/n represents light coupled to the medium. As such, the velocity addition expression given by eqn (2.24) can be applied:

$$v'' = \frac{v+v'}{1+vv'/c^2}. \tag{2.24}$$

Using $v'' = w$ (speed of light relative to the lab), $v' = c/n$ (speed of light relative to stationary water), and v = speed of flowing water with respect to the lab yields

$$w = \frac{c/n + v}{1+v/cn}. \tag{2}$$

The difference between this speed and c/n is

$$w - c/n = \frac{c/n + v - (c/n)(1 + v/cn)}{1 + v/cn} \tag{3}$$

or

$$\boxed{w = c/n \pm \frac{v\left(1-n^{-2}\right)}{1 + v/cn}.} \tag{4}$$

For small values of v, eqn (4) reduces to the Fizeau result given in eqn (1): $w = c/n \pm v(1-n^{-2})$. The ± has been added to take care of the direction of water flow, namely, parallel versus anti-parallel to the wave vector of the light.

Exercise: How does the ± arise in the arrangement shown in Fig. 1? Make a sketch that illustrates why and how the interference is affected by the direction of water flow.

2.4. Lorentz Boost: Active Transformation

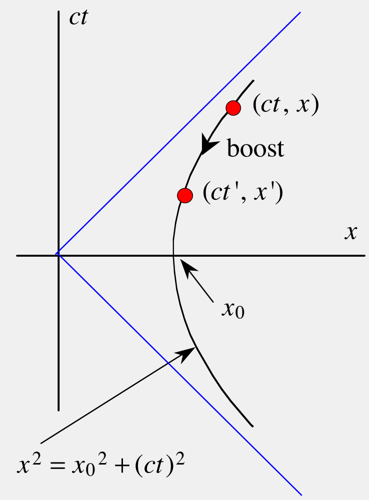


Figure 2.7. This S^2 hyperbola serves as a world line, starting at the bottom of the curve. A boost can move (ct, x) to (ct', x') anywhere on the hyperbola.

The Lorentz boost is now carried out using what is referred to as an active transformation. Referring to the 2D case shown in Fig. 2.7, this transformation acts on a point (ct, x) and moves it to a new location (ct', x'). The new location is determined uniquely by the value of the boost velocity. No set of primed axes is needed, as it is the point that undergoes transformation. This active version of the Lorentz boost complements its passive counterpart that serves as the basis for Minkowski diagrams.

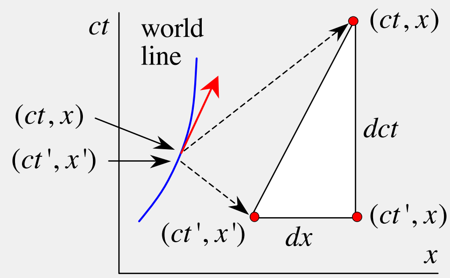
The value of $S^2 = (ct)^2 - x^2$ for some original point, say (ct, x) in Fig. 2.7, is unchanged by a boost. Notice that S^2 has a negative value $(-x_0{}^2)$ for the hyperbola used in the figure. Any point on the hyperbola can be reached through a judicious choice of v, which can be positive or negative. In short, the complete hyperbola is accessible through variation of v.

A continuous change of v along the hyperbola corresponds to acceleration, say, of a point whose world line is the hyperbola. When this acceleration is viewed from the inertial frame indicated in Fig. 2.7, we see that it varies. Coming from below (time always increases), the acceleration increases steadily, reaching a maximum at $x = x_0$ and vanishing as $x \to \infty$. However, the acceleration is seen quite differently from a series of rest frames, spaced infinitesimally from one another, that follow the world line given by the hyperbola. This is discussed later at some length.

In the beginning of this chapter, we saw that in Galilean relativity acceleration is the same in all inertial frames. This apparently is not the case here – or is it! We will see that relativistic acceleration is a four-vector – the Minkowski space analog of the three-acceleration in the Cartesian system used by Galileo. The four-vector (not the column vector that represents it) is invariant, whereas its space part is not.

A world line, like any curve in spacetime, is described by how the spacetime coordinates evolve as some parameter is varied, and the S^2 hyperbola is no exception. The natural parameter is proper time τ, as this has the same value everywhere on the hyperbola. Consequently, it can define four-vectors, as these do not vary from one inertial frame to another.

Referring to the sketch on the right, the difference between two spacetime points that lie near one another, say (ct, x) and (ct', x'), is expressed in terms of differential elements dct and dx. The (red) vector that lies tangent to the (blue) world line is obtained by dividing the vector that connects the points by something that shall be referred to, for lack of a better term, as "the distance between the points." Keep in mind that this "distance" is *not* equal to $(dct)^2 + (dx)^2$, as one might infer from the white triangle. Rather, it is $(dct)^2 - (dx)^2$, because we are working in Minkowski space. This point is clear in hindsight though easy to overlook because we are used to dealing with Euclidean space, and on occasion curved space, for example, the surface of a sphere. Given that $(dct)^2 - (dx)^2 - (dy)^2 - (dz)^2 = (ds)^2 = (dc\tau)^2$, the tangent vector to a timelike curve is seen to be the relativistic velocity:



Note that $(ds)^2 = (dct)^2 - (dx)^2$. It is not equal to $(dct)^2 + (dx)^2$. The red arrow denotes the tangent vector.

$$U^\nu = \begin{pmatrix} \dfrac{dct}{d\tau} \\ \dfrac{dx}{d\tau} \\ \dfrac{dy}{d\tau} \\ \dfrac{dz}{d\tau} \end{pmatrix} = \gamma \begin{pmatrix} c \\ \dfrac{dx}{dt} \\ \dfrac{dy}{dt} \\ \dfrac{dz}{dt} \end{pmatrix}, \tag{2.25}$$

whose four-vector components transform covariantly.

The four-vector itself is the same in all inertial frames, though its representation in terms of components changes from one inertial frame to another. Multiplication by mass, m, gives another four-vector – one whose zeroth component is γmc, which is equal to $\gamma mc^2 / c = E / c$. This four-vector is referred to as the energy-momentum four-vector.

Referring to Fig. 2.7, suppose a particle (or observer) follows the world line given by the S^2 hyperbola, starting at the bottom of the curve and moving upward with increasing time. In the present section, including Example 2.3, we will get a glimpse of how this works. We will find that the proper acceleration – the acceleration experienced by an observer who remains motionless in her rest frame throughout the journey – is constant everywhere on the world line. This will be dealt with in significantly more detail in Section 3.6, but right now we shall see how some important algebra falls neatly into place.

The boost expressions, $ct' = \gamma(ct - \beta x)$ and $x' = \gamma(x - \beta ct)$, are easily organized into compact, suggestive forms by adding and subtracting them:

$$\begin{aligned} ct' + x' &= \gamma\left(ct - \beta x\right) + \gamma\left(x - \beta ct\right) = \gamma(1-\beta)\left(ct + x\right) \\ &= \sqrt{\frac{1-\beta}{1+\beta}}\,\left(ct + x\right) \end{aligned} \tag{2.26}$$

$$\begin{aligned} ct' - x' &= \gamma\left(ct - \beta x\right) - \gamma\left(x - \beta ct\right) = \gamma(1+\beta)\left(ct - x\right) \\ &= \sqrt{\frac{1+\beta}{1-\beta}}\,\left(ct - x\right). \end{aligned} \tag{2.27}$$

The above expressions clearly satisfy $(ct'+x')(ct'-x') = (ct')^2 - (x')^2 = (ct)^2 - x^2$. Furthermore, the forms given by eqns (2.26) and (2.27) invite the definition

$$e^{\phi} = \sqrt{\frac{1+\beta}{1-\beta}}\,. \tag{2.28}$$

This parameterization is intuitive, as e^{ϕ} and $e^{-\phi}$ fit well with eqns (2.26) and (2.27):

$$(ct'+x')(ct'-x') = (ct')^2 - (x')^2 = \cancel{e^{-\phi}}(ct+x)\cancel{e^{\phi}}(ct-x) = (ct)^2 - x^2\,. \tag{2.29}$$

The parameter ϕ is referred to as the *rapidity*. Notice that the exponential form e^{ϕ} enables successive boosts along the same direction to be obtained trivially: $e^{\phi_1 + \phi_2 + \phi_3 \cdots}$. It is easy and useful to express the Lorentz boost in terms of the exponentials e^{ϕ} and $e^{-\phi}$, or, equivalently, their hyperbolic counterparts: $\cosh\phi$, $\sinh\phi$, and $\tanh\phi$. To do this, write out $(e^{\phi} + e^{-\phi})^2$ using eqn (2.31):

$$(e^{\phi} + e^{-\phi})^2 = 4\cosh^2\phi = \left(\sqrt{\frac{1+\beta}{1-\beta}} + \sqrt{\frac{1-\beta}{1+\beta}}\right)^2 = \frac{4}{1-\beta^2} = 4\gamma^2 \tag{2.30}$$

We see that $\gamma = \cosh\phi$, and minor fiddling yields $\beta = \tanh\phi$ and $\beta\gamma = \sinh\phi$.[10] These expressions will prove extremely useful. Introducing them to a familiar boost gives

[10] $\cosh^2\phi - \sinh^2\phi = 1$. Thus, $\sinh^2\phi = \gamma^2 - 1 = \dfrac{1-1+\beta^2}{1-\beta^2} = (\beta\gamma)^2$ and $\tanh^2\phi = \dfrac{\sinh^2\phi}{\cosh^2\phi} = \beta^2$.

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \cosh\phi & -\sinh\phi & 0 & 0 \\ -\sinh\phi & \cosh\phi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} \tag{2.31}$$

This hyperbolic form combines nicely with 3D rotation, as discussed in Section 6.[11] Indeed, we see that the natural representation of Lorentz boosts is hyperbolic. This is a direct consequence of working in Minkowski space, where the invariant 2D quantity is $(ct)^2 - x^2$, as opposed to $x^2 + y^2$ in Euclidean space. A 2D rotation in Euclidean space is carried out using $\cos\theta$ and $\sin\theta$, whereas in Minkowski space the analogous operation uses $\cosh\phi$ and $\sinh\phi$. The rapidity parameter ϕ (actually, $\cosh\phi$ and $\sinh\phi$) is generally more convenient than γ and β when it comes to carrying out calculations, so we shall use it frequently.

So far, we have considered boosts in the same direction: $e^{\phi_1+\phi_2+\phi_3\cdots}$. What about boosts in different directions, for example, in orthogonal directions? In 1926, the English physicist Llewellyn Hilleth Thomas (1903-1992) showed that sequential boosts in different directions result not simply in another boost, but a boost plus a rotation. He was a graduate student at the University of Cambridge at the time, though on a one-year fellowship at Niels Bohr's institute in Copenhagen. This effect is referred to as Thomas rotation. Eugene Wigner derived it from group theoretical considerations much later, in 1939, so it is sometimes referred to as Thomas-Wigner rotation. How and why people's names get attached to equations and the like is interesting.[12] It can be used to work out a relativistic effect referred to as Thomas precession.


Llewellyn Hilleth Thomas

[11] You might find it helpful to read through Section 6 at this point. It is short. You can get through it in an hour or less.

[12] The algebras of the Pauli and Dirac and matrices were introduced a half-century before quantum mechanics by the English mathematician William Kingdon Clifford. Oliver Heaviside derived the Lorentz force law several years before Hendrik Antoon Lorentz's derivation. Harold Jeffreys worked out the WKB approximation several years earlier than did Gregor Wentzel, Hendrik Kramers, and Leon Brillouin. Ehrenberg and Siday introduced the Aharonov-Bohm effect a decade before Aharonov and Bohm. Henri Poincaré introduced the Lorentz group. And so on.

Before proceeding, it is worth noting that the light lines constitute a natural set of orthogonal axes. The idea is shown in Fig. 2.8, where Greek letters ξ and η (*xi* and *eta*) label the light-line axes. Hyperbolas contained within the ξ and η axes (*e.g*., the black one) have as asymptotes the ξ and η axes. An infinite number of such hyperbolas satisfy $(ct)^2 - x^2 = constant$, that is, for a continuous range of values for the *constant*. Straightforward geometry converts (ct, x) to its (ξ, η) counterpart. Namely, projecting ct and x alternately onto the η and ξ axes,[13] yields

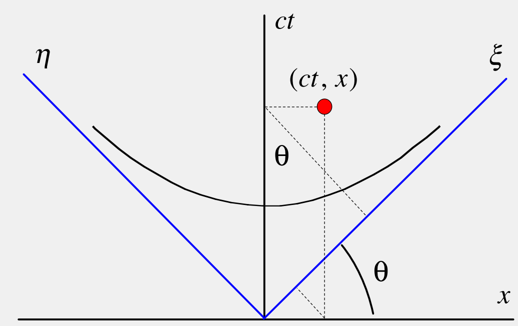


Figure 2.8. The blue light lines are asymptotes of the hyperbola. Coordinates (ct, x) are projected onto the (ξ, η) axes.

$$\xi = \frac{x + ct}{\sqrt{2}} \text{ and } \eta = \frac{ct - x}{\sqrt{2}} . \tag{2.32}$$

Exercise: Start with the hyperbola and axes used in Fig. 2.7 and carry out projections analogous to those that yielded eqn (2.32). Do you get the same result or a slightly different one? Explain why your result is the same or different.

Example 2.3. Constant Acceleration

We shall end this section with the example of constant acceleration that was mentioned at the beginning of the section. Consider the world line given by the hyperbola in Figure 1. An observer commences travel along this world line with a large positive value of x at the lowest part of the hyperbola. In fact, think of the observer as starting at a value of x that is much larger than can be shown in the figure. Her speed is close to c. She moves toward the left and slows with time until finally she reaches the turning point at $t = 0$. From there she moves to the right and speeds up, eventually reaching a speed close to c. She undergoes acceleration throughout the trip and notes that, in fact, this acceleration is constant.

Notice that it is the observer moving on her world line who finds that her acceleration is constant. Certainly a person watching her from an inertial frame would not measure a constant acceleration. Were this latter person to observe an object undergoing constant

---

[13] The geometry in Fig. 2.8 yields $\xi = (ct)\sin\theta + x\cos\theta$ and $\eta = (ct)\sin(90° - \theta) - x\cos(90° - \theta)$. Using $\theta = 45°$ yields $\xi = (ct + x)/\sqrt{2}$ and $\eta = (ct - x)/\sqrt{2}$, which is eqn (2.32).

acceleration, $\ddot{x}$, for a sufficiently long time, the object's speed would eventually exceed the speed of light!

Let us now establish the reason that the acceleration experienced by the observer is constant. First, however, let us go back to the section in the very beginning of this chapter entitled: Galilean Transformation. There we saw that the physical vector acceleration is the same in every inertial frame. This was for the Galilean case, and in general relativistic acceleration is more complicated. Nonetheless, something similar happens here.

We shall start by placing a point (red dot) anywhere on the hyperbola. An instantaneous (local) inertial frame is assigned to the observer when she is located at the red dot. She is stationary at its origin, so it is her rest frame. Of course, the observer is undergoing acceleration throughout the entire trip on the hyperbola. Consequently, the use of an instantaneous inertial frame is with the understanding that this frame changes continuously throughout the journey in accord with v changing continuously.

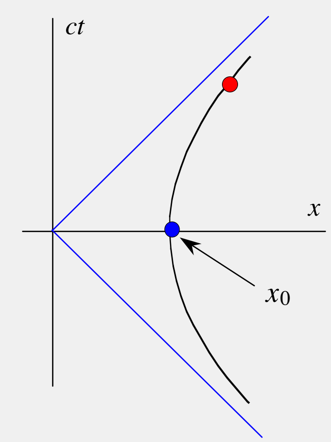


Figure 1. The hyperbola is a world line. A particle at the red position is boosted to the blue position.

A Lorentz boost now moves the system to a new instantaneous inertial frame, whose location is indicated using a blue dot. It can be anywhere on the hyperbola. We shall place the blue dot on the x-axis.

The observer finds that the blue frame is also her rest frame. After all, her world line is vertical at $t = 0$. The location of the red point was arbitrary, so it must be the case that the observer remains fixed to the origin of her reference frame everywhere on her hyperbolic world line. Consequently, she experiences proper time and proper acceleration and she measures proper lengths. An active boost can move her to another inertial frame, but it cannot change her proper acceleration or any proper length she might measure. Thus, the observer experiences constant acceleration throughout the entire journey.

An especially interesting feature is the following. Suppose a pulse of light is launched in the $+x$-direction at $t = 0$, *i.e.*, from the origin in Figure 1. It is a distance x_0 from the observer, who is at the blue point at $t = 0$. The observer uses her lattice of clocks to measure the proper distance x_0 . Later the observer is at the red position. She again uses her lattice of clocks to locate the light pulse. She finds that it is still a distance x_0 behind her. It has not caught up. Nor will it ever catch up, as the location of the red dot is arbitrary. In other words, the light pulse never catches the observer, who experiences constant proper acceleration along the hyperbola.

This example, in fact all of Section 2.4, has been hurried. The idea was to introduce the hyperbolic nature of Lorentz transformations and how it relates to acceleration. We shall return to acceleration in Section 3.4, where it is examined in more detail.

Example 2.4. Doppler Effect (Bibek Presentation)

The Doppler effect arises frequently in spectroscopic studies of gas phase atoms, small molecules, and small radicals. It plays a special role with ions because they can be accelerated to high velocities. Moreover, because the velocity of an ion can be controlled (and modulated) through adjustment of the electric potential difference through which the ion travels, it is possible to use the Doppler effect to tune an ionized species' absorption lines into coincidence with a fixed laser frequency.

With molecular beams of neutral species, it is possible to minimize the degree of spectral shift and broadening that the Doppler effect incurs, for example, by arranging experimental conditions such that the velocity of the centermost part of the molecular beam is perpendicular to the $\vec{k}$-vector of the incident radiation. Sometimes this is referred to as Doppler-free spectroscopy. However, this is not correct. The Doppler effect is not eliminated. It is simply small because of the geometrical arrangement. Were one to spectroscopically probe the molecular beam head-on rather than from the side (arrange for the laser and molecular beams to counter-propagate), the Doppler shift would manifest, and it can be significant.

On the other hand, 2-photon excitation using counter-propagating laser beams (wave vectors $\vec{k}$ and $-\vec{k}$) can be used to carry out Doppler-free spectroscopy. The non-relativistic Doppler shifts for the $\vec{k}$ and $-\vec{k}$ contributions to the 2-photon excitation rate add to zero, regardless of the size of the velocity component in the direction of $\vec{k}$.

Frequency Shift

Referring to the right hand side of Fig. 1, an electromagnetic wave of frequency ω_0 is emitted in a primed reference frame at its origin. This frame is in motion (red arrow) relative to an observer at rest at the origin of an unprimed frame. The frames coincide at $t = 0$ and $t' = 0$. In the moving frame, a cycle of the electromagnetic wave starts at $t' = 0$ and ends at the completion of one period: $t' = \tau_0 = 2\pi / \omega_0$.

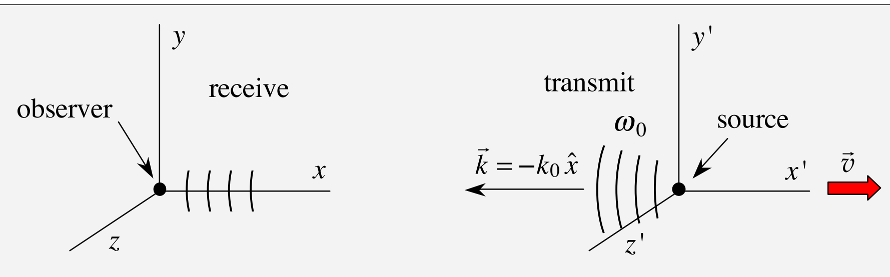


Figure 1. A source emits photons of frequency ω_0. Because the source-to-observer distance increases (red arrow), the observer senses a lower frequency than ω_0. In addition, time dilation adds a factor of γ.

According to the person who manages the source of radiation (and resides in the primed frame), the distance that light emitted at τ_0 travels to reach the observer is the relative speed times the cycle period: $v\tau_0$. However, when the observer in the unprimed (stationary) frame measures this distance she finds $v\tau = v\gamma\tau_0$ due to time dilation. In the observer's frame, the time at which the end of the cycle arrives (τ) contains two contributions. One is the observer's measurement of the time when the end of the cycle left the source region, $\gamma\tau_0$. The other is the distance traveled by the source when the end of the cycle leaves the source region divided by the speed of light. Thus, we have the relationship between the periods and frequencies of the emitted and received photons:

$$\tau = \gamma\tau_0 + v\gamma\tau_0 / c \tag{1}$$

$$= \gamma(1+\beta)\tau_0 \tag{2}$$

$$= \tau_0\sqrt{\frac{1+\beta}{1-\beta}} \,. \tag{3}$$

For the corresponding frequencies we have

$$\omega = \omega_0\sqrt{\frac{1-\beta}{1+\beta}} \,. \tag{4}$$

For small β values this yields an expression you surely have encountered elsewhere:

$$\omega \approx \omega_0(1-\beta) \,. \tag{5}$$

This result, which is valid throughout the non-relativistic regime, is referred to as the longitudinal Doppler effect, or (usually) simply the Doppler effect. Had we considered the source approaching the observer, eqn (5) would have read: $\omega \approx \omega_0(1+\beta)$. Some realistic numbers will be given later for typical Doppler shifts, and to elucidate the relativistic versus non-relativistic regimes of eqn (4).

Transverse Doppler Shift

The interesting case of the Doppler shift that is present even though the velocity is perpendicular to the direction of observation is handled most easily by Lorentz transformation of the wave four-vector $(\omega/c, \vec{k})$. As in Fig. 1, the light source is fixed at the origin of the primed frame, and the observer is at rest at the origin of the unprimed frame. Let us begin by applying a Lorentz transformation to the case depicted in Fig. 1, in which $\vec{v}$ is parallel to the direction of observation. This is simply an alternate way to obtain the result given by eqn (4).

$$\begin{pmatrix} \omega / c \\ k_x \\ k_y \\ k_z \end{pmatrix} = \begin{pmatrix} \gamma & \pm\gamma\beta & 0 & 0 \\ \pm\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \omega_0 / c \\ -k_0 \\ 0 \\ 0 \end{pmatrix} \tag{6}$$

where $k_x' = -k_0$ and $\omega' = \omega_0$. The $\pm$ deals with the two orientations of $\vec{v}$, *i.e.*, source and observer moving toward versus away from one another. Equation (6) yields

$$\omega = \gamma\omega_0 \pm \gamma\beta\, k_0 c\,. \tag{7}$$

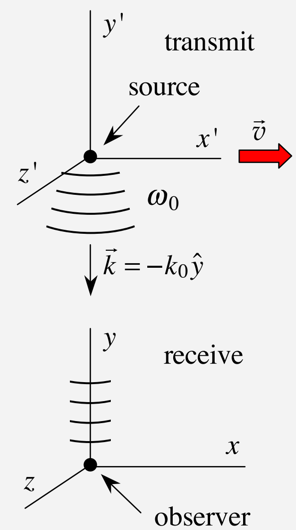
In the primed frame we have $\omega_0 = k_0 c$. Thus, eqn (7) becomes

$$\omega = \omega_0 \gamma (1 \pm \beta)\,. \tag{8}$$

Again, this is the longitudinal Doppler effect given by eqn (4), albeit this time with $\pm$ included for the two velocity orientations.

Now consider the case in which the light source passes the observer with $\vec{v}$ and $\vec{k}$ perpendicular to one another, as indicated in Fig. 2. This gives rise to what is called the transverse Doppler effect. It vanishes in the non-relativistic limit. The matrix representation of the Lorentz transformation is again used to calculate the frequency detected by the observer in the unprimed frame. You can see from Fig. 2 that the β that appears in the Doppler shift given by eqn (8) vanishes for this configuration. However, time dilation is carried by γ and this persists. Applying the Lorentz transformation when the source is directly overhead, *i.e.*, $\vec{k} = -k_0\,\hat{y}$ yields

Figure 2. The radiation source moves in a direction that is perpendicular to that of observation. This eliminates the non-relativistic β part of the Doppler effect, but not the relativistic γ part that arises from time dilation.



$$\begin{pmatrix} \omega / c \\ k_x \\ k_y \\ k_z \end{pmatrix} = \begin{pmatrix} \gamma & \pm\gamma\beta & 0 & 0 \\ \pm\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \omega_0 / c \\ 0 \\ -k_0 \\ 0 \end{pmatrix}. \tag{9}$$

This gives the remarkably simple result:

$$\omega = \gamma\omega_0\,. \tag{10}$$

This is the transverse Doppler shift. It is due to time dilation. There is no contribution due to the relative motion of the source and observer in the direction of observation. The genera case considers arbitrary alignment and orientation of $\vec{k}$ and $\vec{v}$.

Estimates

Numerical estimates related to laboratory experiments and celestial observations improve one's intuition. Let us start with atomic hydrogen, which is frequently detected via the Lyman-α transition at 121.6 nm (≈ 82 240 cm^{-1}). Ultraviolet photodissociation can be used to create H-atoms with speeds as high as ~ 3×10^{6} cm s^{-1}, in which case we can take $\beta = 10^{-4}$. This lies solidly in the non-relativistic regime given by eqn (5). That is, $\beta = 10^{-4}$ gives $\gamma = 1.000000005$. Using eqn (5), the resulting Doppler shift is $\beta\omega_0 = 8.224$ cm^{-1}, which is large and easily measured.

Next, consider the Doppler shift that can arise when a transition takes place in a light molecular ion after it has been accelerated through a significant potential difference. In our laboratory, it is not unreasonable to use voltages as large as 100 kV, though one must be careful around such high voltages. Suppose that N_2^+ is accelerated to 100 kV and then examined spectroscopically using radiation at 25 000 cm^{-1}. From the above equations, at 100 kV the system is, for the present purposes, still in the non-relativistic regime: $\beta = 0.0276$ and $\gamma = 1.0004$. The non-relativistic Doppler shift is large: $\beta\omega_0 = 690$ cm^{-1}. The center wavelength has moved from 400 nm to 411.6 nm.

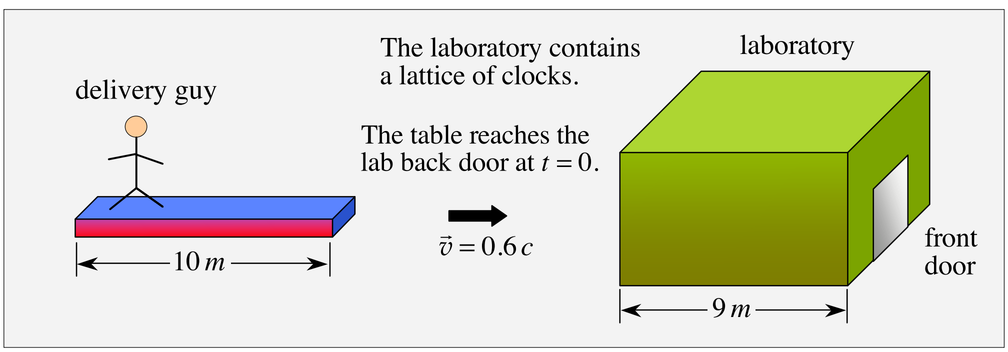
Finally, consider a highly relativistic species, say $\beta = 0.999$. In this case, $\gamma = 22.36$ and the lower of the two frequencies given by eqn (8) is $\omega = \gamma(1-\beta)\omega_0 = 0.0226\,\omega_0$. The longitudinal Doppler shift is enormous. For example, Lyman-α radiation is Doppler shifted into the infrared (1859 cm^{-1}). The higher of the two frequencies given by eqn (8) gives an equally dramatic shift: $\omega = \gamma(1+\beta)\omega_0 = 1{,}837{,}050$ cm^{-1}, corresponding to $\lambda = 5.44$ nm. This is essentially a factor of two larger than the transverse Doppler shift, again an enormous effect. Such highly relativistic situations arise with light arriving on earth from distant parts of the universe. Such observations have nothing to do with chemical physics, but they are interesting nonetheless.

Example 2.5. Relativistic Laser Table (Parmeet Presentation)

An experimenter ordered a large laser table from a Silicon Valley startup company that advertised prompt delivery and a high quality product. The experimenter was in a rush to place the order so she did not check references. She needed the table "yesterday."

The delivery turned out to be the opposite of prompt. Several months passed with one excuse after another, rendering the quality of the product moot. As you might imagine, this infuriated the experimenter. The company finally completed its manufacture of the table. They were desperate to keep the experimenter from informing everyone she would meet at a large, upcoming scientific conference of her difficulties. Thus, they decided to make amends by enlisting a delivery guy whose specialty was speedy service. The delivery guy announced that he would pull out all stops and transport the table at near the speed of light.

The delivery day finally arrived. The table approached the laboratory at a velocity of $0.6c$, as indicated in the sketch below. In their respective rest frames the table was 10 meters long, whereas the laboratory was only 9 meters long. The experimenter measured the rapidly moving table with a synchronized lattice of clocks and found that there was a brief period during which it was indeed completely inside the laboratory, though it did not slow and went zooming out the front door. For example, it proved possible to close the back and front doors simultaneously (but for just an instant) with the table completely inside the lab during this brief interval, before opening them wide again. This proved conclusively that the table fitted inside the laboratory, if only for a fleeting moment.



However, the delivery guy was traveling atop the table, and he disagreed, claiming that the table simply did not fit into the lab at any point in time. He claimed that this is the reason he did not stop, or return to the lab for a second delivery attempt. Instead, he continued through the open front door and returned to the company. The delivery guy had his own properly synchronized lattice of clocks, and he was adamant.

Let us try to figure out what happened.

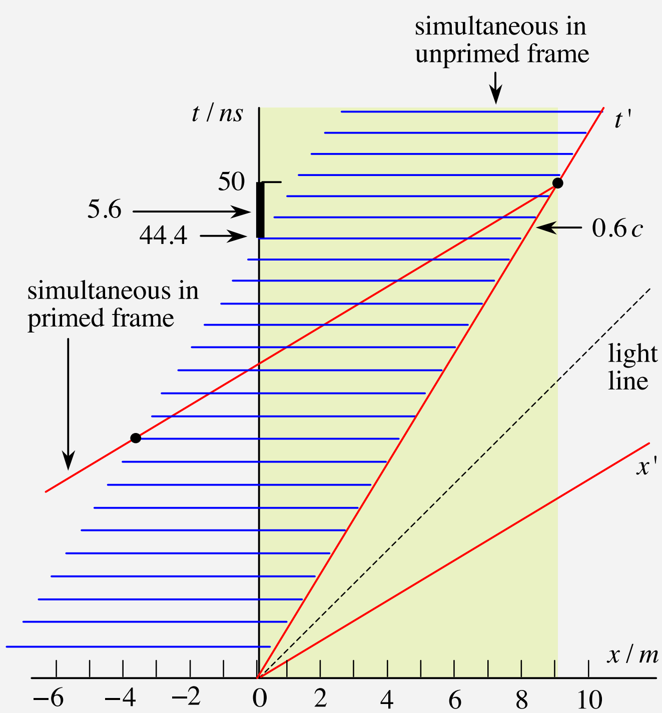
Suppose we take the laboratory measurements that were carried out using the experimenter's lattice of clocks at face value. In this case, there is no doubt that the rapidly moving table fitted inside the lab. Its length was consistent with the calculated contraction. Namely, a speed of $0.6c$ gives $\gamma = 1.25$, and length contraction makes the table only 8 meters long. Clearly it fitted inside the lab.

However, the delivery person was no slouch at his trade. He was correct in his finding that the length of the laboratory was only 7.2 meters long due to its length contraction, whereas the table was 10 meters long in its rest frame.

The issue is one of simultaneity. Looking at it from the perspective of the delivery guy, the table entered the laboratory through the open back door and went toward the open front door. Before it reached the open front door, this door shut and reopened quickly. The table then passed through the open front door. At this point the rear of the table had not yet reached the back door. When it did, it passed into the laboratory, and the rear door shut and reopened quickly. Thus, the front and back doors closed and opened just as the experimenter said. However, in the reference frame of the delivery guy, these were not simultaneous events.

The mystery is solved. To be more quantitative, however, let us now examine the Minkowski diagram for this system.

Minkowski diagram: The 9-meter length of the lab is represented in the unprimed frame with light green shading. The length-contracted table is represented with horizontal blue lines that are parallel to horizontal lines of simultaneity, the density of which is infinite. In the primed frame, lines parallel to the x' axis are lines of simultaneity. In the unprimed frame, the table fits into the 9-meter length of the lab, whereas in the primed frame it never fits into the lab. The two doors open and close quickly and simultaneously in the unprimed frame. In the primed frame they do this sequentially, first the front door, then the rear door.



The Minkowski diagram enables us to see right away what happened. In the experimenter (unprimed) frame the table moves with a speed of $0.6\ c$ ($\gamma = 1.25$). The front and back ends of the table (whose length is indicated by blue lines) are 8 meters apart due to length contraction. The front of the table arrives at the back door at $t = 0$. The table passes into the lab, and at $t = 44.4$ ns it is completely within the lab. It remains within the lab for a period of 5.6 ns (heavy black line on the t-axis). It begins exiting the lab at $t = 50$ ns.

The delivery guy finds that the table enters the lab at $t' = 0$, but he finds that the lab is only 7.2 meters long according to the x' axis on his lattice of clocks. When the table reaches the front door of the lab (upper right black dot), he finds, again using his lattice of clocks, that the rear of the table (lower left black dot) has not yet entered the lab. The closing and opening of the doors makes sense, though we shall avoid a discussion of relativistic doors.

Tempers cooled after a while, and the experimenter and delivery guy decided to try again. This time the front door would be kept shut. The front door would be made of a proprietary "super-material" that is not incompressible, but close to it. The experimenter's idea was that the table would stop when it hit the front door. The rear door (which was also made of the super-material) would be quickly shut right before the table collided with the front door, ensuring containment of the table within the lab.

On the day of the second delivery attempt, the various clocks were synchronized and the table zoomed into the lab. It stopped when it hit the front door. This required world-class deceleration, but this is no problem in a gedanken experiment. According to the experimenter, the rear door shut before the table hit the front door, thereby trapping the table inside the lab. However, once the table experienced its rapid deceleration, its length was no longer contracted. It tried to be 10 meters long, and therefore it was crumpled between the front and back doors. The crumpling began at the front of the table and propagated toward the rear of the table. Laser tables may look strong, but the body is made of metal honeycomb, the same stuff used to make airplane fuselages, so it is easily crumpled. Anyway, the table was destroyed. So much for a high quality product!

The delivery guy used his lattice of clocks to paint a different picture. He agreed on the event in which the table hit the front door, but he found that the rear of the table had not yet entered the lab. According to him, upon the table's impact with the front door, it underwent a compression that began at its front and propagated toward its rear.

The delivery guy pointed out that it is impossible for a table to be 100% incompressible, as this would require that the speed of sound is infinite. The experimenter agreed. She remembered from Math Methods that this could not happen, so she acknowledged that the table would crumple in a manner that resembles propagation of some kind. The crumpling would be such that the rear of the table was inside the lab when the rear door closed.

After all was said and done, the experimenter and delivery guy agreed on the end result – the laser is inside the lab, albeit destroyed.

Exercise: Construct a Minkowski diagram that illustrates this scenario.

Example 2.6. Relativistic Slab and Platform (Bailey Presentation)

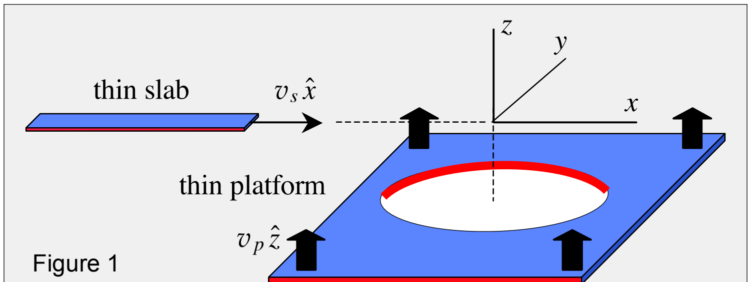


Figure 1

A thin slab travels at a large velocity $v_s\hat{x}$ relative to the stationary laboratory frame. An equally thin platform with a circular hole in its center moves in the z-direction with velocity $v_p\hat{z}$ relative to the laboratory frame. The thin slab is one meter long in its rest frame, and the diameter of the hole is one meter in the platform's rest frame. With this arrangement, were $v_s\hat{x}$ small, the slab and platform would collide, and the slab would not pass through the circular opening as the platform moves upward.

For the case at hand of large $v_s\hat{x}$, an observer in the laboratory finds that when the platform reaches the xy-plane, the center of the slab is at $x = y = 0$, and ever so slightly above the platform. The velocity $v_s\hat{x}$ is large enough that the length of the slab is contracted significantly in the laboratory frame. This enables the platform to pass upward, past the slab, without touching it. On the other hand, from the perspective of the slab's rest frame it is the platform whose length is contracted significantly. Indeed, as viewed from the slab's rest frame, the length of the opening is quite a bit smaller than the length of the slab.

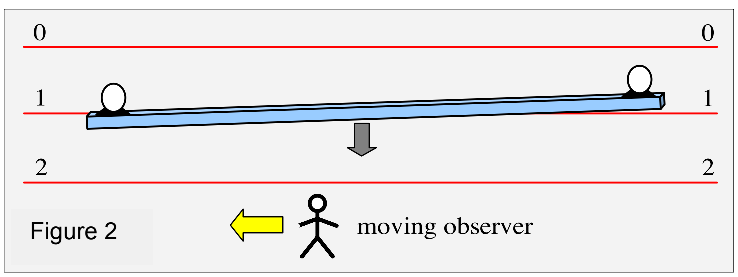


Figure 2

How can these irrefutable facts and the two perfectly legitimate perspectives be reconciled? To understand better this situation, go back and reread Examples 1.1 and 1.2. The most relevant figure is reproduced on the right. Reconciliation is achieved by taking lack of simultaneity into account.

From the perspective of the slab, the hole in the platform is contracted. However, the platform is also rotated. Using Fig. 2 as a guide, we find that the platform is rotated counterclockwise. This enables the slab to pass through the opening, which has been contracted in the x-direction. Though going from Fig. 2 to Fig. 1 requires careful bookkeeping, with sufficient patience you will find that everything falls into place. Bailey used a few equations, but I think the pictures create a mental image that makes it easier to remember what is going on

3. Four-Velocity, Four-Momentum, and the Like

When Gauss was nineteen his mother asked his mathematical friend Wolfgang Bolyai whether Gauss would ever amount to anything. When Bolyai exclaimed "the greatest mathematician in Europe!" she became distraught and burst into tears.

In this section we shall examine some kinematical and dynamical consequences of the boost. To begin, four-velocity is defined, and from there four-momentum is obtained in a few steps. This approach is sensible and useful. Though the introduction of four-velocity in this manner might seem a bit *ad hoc*, it is on solid ground. Regardless, it is followed by a clever derivation of the four-momentum [5]. A modest amount of algebra then yields the momentum-energy relationship that is central to relativistic dynamics.

To distinguish three-vectors and four-vectors, Latin subscripts and superscripts are used with three-vectors, whereas Greek subscripts and superscripts are used with four-vectors. We shall begin with a light-hearted summary of results derived at the beginning of Section 1, starting with the boost for velocity $\vec{v}$ parallel to the x-axis. The transformation from the unprimed to the primed frame, with O moving in the $-x$ direction relative to O', is given by

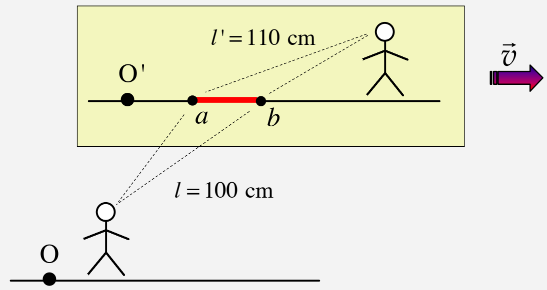
$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}. \tag{3.1}$$

Alternatively, an observer in O sees O' moving with a velocity that points in the $+x$ direction. In this case, the counterpart to eqn (3.1) is

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \gamma & \beta\gamma & 0 & 0 \\ \beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix}. \tag{3.2}$$

The situation is illustrated in Fig. 3.1. An observer in frame O records a length l between the points a and b at the ends of a rod that is stationary in O'. In other words, the length l is that of a rod that is moving relative to the observer in O. However, were the observer to place herself in the moving frame and measure the length there, she would find that it is larger than l by a factor of γ, as indicated in Fig. 3.1. This phenomenon, in which a moving object appears to be shorter than when it is stationary, is referred to as length contraction or Lorentz contraction. It was derived in Section 1 in the subsection entitled: Length Contraction along the Direction of Motion.

Figure 3.1. The O' frame moves with velocity $\vec{v}$ relative to O. Points a and b are stationary in O'. An observer in O measures a distance of 100 cm between a and b. An observer in O' moves with a and b and records 110 cm.



The square matrices in eqns (3.1) and (3.2) differ only in the sign of β, which reflects the velocity direction. In cases like the above, assignment of motion to the x-axis is possible because the x-axis can be defined such that it is parallel to $\vec{v}$. On occasion one must deal with more tedious expressions. The transformation analogous to eqn (3.1) (for $\vec{v}$ in an arbitrary direction) is given below. You are welcome to work out the details if you so choose. We shall avoid such messy expressions.

$$\Lambda = \begin{pmatrix} \gamma & -\beta_x\gamma & -\beta_y\gamma & -\beta_z\gamma \\ -\beta_x\gamma & 1+\beta_x{}^2/\beta^2(\gamma-1) & 1+\beta_x\beta_y/\beta^2(\gamma-1) & 1+\beta_x\beta_z/\beta^2(\gamma-1) \\ -\beta_y\gamma & 1+\beta_x\beta_y/\beta^2(\gamma-1) & 1+\beta_y{}^2/\beta^2(\gamma-1) & 1+\beta_y\beta_z/\beta^2(\gamma-1) \\ -\beta_z\gamma & 1+\beta_x\beta_z/\beta^2(\gamma-1) & 1+\beta_y\beta_z/\beta^2(\gamma-1) & 1+\beta_z{}^2/\beta^2(\gamma-1) \end{pmatrix}. \tag{3.3}$$

Clearly, using matrix multiplication to compound successive transformations can get unruly, and eqn (3.3) does not even include 3D rotation, which is part of the Lorentz group. It is not necessary in our introductory treatment. Were it needed, it would be carried out trivially using a computer.

3.1. Invariants

The squared interval will now be used to obtain an expression for a velocity that transforms as a four-vector. The strategy that underlies the use of four-vectors and other Lorentz covariant (and invariant) quantities (scalars, tensors of rank 2 and higher) is straightforward. By constructing models from ingredients that are Lorentz covariant and invariant, the models satisfy the requirements of special relativity.

Earlier it was pointed out that the scalar $S^2 = (ct)^2 - x^2 - y^2 - z^2$ is a conserved quantity: the squared interval between (ct, x, y, z) and the origin $(0,0,0,0)$. The more general expression uses the difference between events, say (ct_1, x_1, y_1, z_1) and (ct_2, x_2, y_2, z_2), yielding $(c(t_2 - t_1), (x_2 - x_1), (y_2 - y_1), (z_2 - z_1))$. Using $c(t_2 - t_1) = cdt$, $x_2 - x_1 = dx$, and so on, the squared interval becomes

$$ds^2 = c^2dt^2 - \left(dx^2 + dy^2 + dz^2\right) = c^2dt^2\left(1 - \frac{dx^2 + dy^2 + dz^2}{c^2dt^2}\right). \tag{3.4}$$

It is understood that dt^2 stands for $(dt)^2$ rather than $d(t^2)$, and likewise for s and the x, y, z coordinates. Note the use of a small s. The d used here can indicate a differential or a difference. Some authors use Δ for the latter, but we will use d, with the distinction between its use as a differential versus a finite difference taken from context. Using the fact that $dx^2 + dy^2 + dz^2 = v^2 dt^2$, eqn (3.4) becomes

$$ds^2 = c^2 dt^2 / \gamma^2 = c^2 d\tau^2 , \qquad (3.5)$$

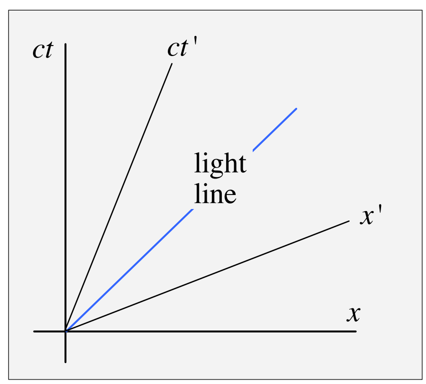
where $d\tau = dt / \gamma$.

3.2. Proper Time

Taking the positive square roots of the terms in eqn (3.5) yields

$$ds = d(c\tau) .$$

Because the scalar ds^2 is invariant, so is $d\tau$. The quantity $\tau = t / \gamma$ is referred to by several names: proper time, local time, and the Lorentz invariant element of time. The proper time between two events is the measured time in the rest frame of the object under consideration. By definition, the object follows a spacetime world line along the time axis. Again, being a Lorentz invariant scalar, the proper time has the same value in all inertial frames.



Material quantities measured in their rest frames are referred to as proper. For example, the proper distance between two events is the distance between them measured in an inertial frame in which the events are simultaneous (imagine lines parallel to the x and x' axes in the sketch). The proper length of an object is the length measured by an observer at rest relative to the object.

A great thing about proper time is that it has the same value in all inertial frames. Together with Minkowski diagrams it is a great help insofar as revealing subtleties of simultaneity and lack thereof. It finds its way into relativistic four-velocity and four-momentum. The following example illustrates its use in a Minkowski diagram to explain yet another of the simultaneity paradoxes.

Example 3.1. The Twin Paradox

Identical twins Delirious and Serious have stumbled on a once-in-a-lifetime opportunity: high-speed space travel on their own personal spaceship for as long a trip as they like. All expenses are covered through a grant to USC from the National Science Foundation under their Transformative Enlistment of Aliens Initiative. This program supports diversity-driven research aimed at contacting aliens from outer space and then carrying out collaborative research with them. The grant focuses on aspects related to traffic control at relativistic speeds and non-surgical slowing of the aging process through time dilation.

Delirious (hereafter referred to as D) declines the offer, choosing instead to remain on earth, presumably to save it. On the other hand, Serious (hereafter referred to as S) is inclined toward outer space, so she accepts the offer. In due course she enters the assigned spaceship and commences travel on a straight-line path, accelerating quickly to a constant speed of $v = 0.866\,c$ relative to earth,[14] which we take to be an inertial reference frame.

After 5 years, S decides that she has seen enough of outer space. The initial excitement passed, and life in a spaceship, in fact, is remarkably boring. She turns the spaceship around quickly and begins the return trip to earth, again at $v = 0.866\,c$. When she arrives on earth S finds to her surprise that D looks quite a bit older. At first she thinks this is a consequence of D's efforts to save the earth, an arduous task if ever there was one. Upon further inspection, however, she finds that D has aged much more rapidly than her: 20 years, in contrast to her own 10 years of aging. D explains that this is due to time dilation. S's clock simply ran more slowly. However, S is not convinced. For example, except for brief periods of acceleration, the spaceship is also an inertial frame. As judged from the inertial frame of the spaceship, it was D's clock that ran slowly during the outbound and inbound parts of the trip. So perhaps D should be the younger twin. S is perplexed.

Apparent Contradiction

Shortly before the flight's departure, S had listened to Stephanie's mini-seminar on special relativity – not a bad idea before embarking on a high-speed journey to outer space. This gives her confidence. She knows she has the intellectual tools needed to resolve the issue. She also knows that D understands special relativity. D not only attended Stephanie's seminar, but she even asked questions. Between the two of them, surely they can get past any apparent contradiction and determine what is going on.

Resolution of the twin paradox provides insight into relationships between measurements of time and space carried out in different inertial frames. Can you figure out before we go any further which twin is younger at the end of the trip? Note: (1) This is not a trick question; their ages indeed differ. (2) The acceleration that takes place in the region where the spaceship turns around is central, but in a way that is subtle. (3) From S's perspective, the time periods over which acceleration takes place during takeoff, turnaround,

[14] Use of 0.866 is convenient. Being $\sqrt{3}/2$, squaring gives ¾ and a Lorentz factor γ of 2.

and landing can be taken as negligible. (4) Because of assumption (3), resolution can be achieved within the framework of special relativity.

As you might have guessed, the culprit is some implicit assumption concerning simultaneity. Consequently, this is not a true paradox (logical inconsistency), but an example of a subtle point that can be understood through careful bookkeeping. Figure 1 illustrates the outbound portion of the trip. From D's perspective, time passes more slowly in S's frame than in hers by a factor of two ($t = \gamma t'$, where $\gamma = 2$ for $v = 0.866\,c$). This dilation is true regardless of whether the spaceship is traveling away from, or toward, the earth.

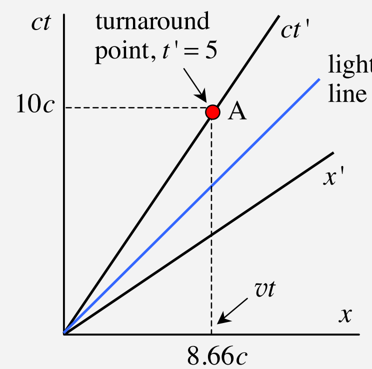


Figure 1. D and S have unprimed and primed frames, respectively. At point A, D's lattice of clocks indicates that the time elapsed on earth is twice the time elapsed in the spaceship (see Fig. 2).

Throughout the round trip, S knew this was the perception from earth, so on this basis she had reason to believe she would be 10 years younger than D when they met again. They were each 25 years old when the trip started, so at the trip's end she would be 35 and D would be 45. And this is exactly what she encountered when arriving back on earth after 10 years of space travel.

Nonetheless, S was skeptical. It seemed to her that such a scenario is not even-handed insofar as its treatment of observers at rest in inertial frames. For example, if it is assumed that D is the one in motion, then D is the younger twin at the end of the journey. At the same time, it was S who, in fact, turned around, not D, so perhaps the act of turning around brought about an effect of some kind. These and other ambiguities cause S to leave open the question of whether their ages differ, and if so, which of them is younger, and by how much.

From Takeoff to Turnaround

There is nothing in need of reconciliation for the first part of the trip: from takeoff to where the spaceship begins the intense acceleration that turns it around at point A. For this part of the trip, D and S interpret their measurements as straightforward manifestations of special relativity. Referring to Fig. 2(a), when D records $t = 10$ years, snapshots are taken from throughout her lattice of clocks, as indicated by red arrows pointing downward. These snapshots record locations and clock settings of S's lattice of clocks.

D finds, using her lattice of clocks, that when $t = 10$ (years) the spaceship is at the turnaround point A, and that S measures $t' = 5$ on the clock she carries with her. On the other hand, when S measures $t' = 5$ on her own clock and then uses her lattice of clocks to record the settings on D's lattice of clocks, S finds that D's clock reads $t = 2.5$. From their understanding of special relativity, D and S agree that the situation is symmetric. Each finds the same time dilation ($\gamma = 2$) for measurements carried out in the other person's inertial reference frame.

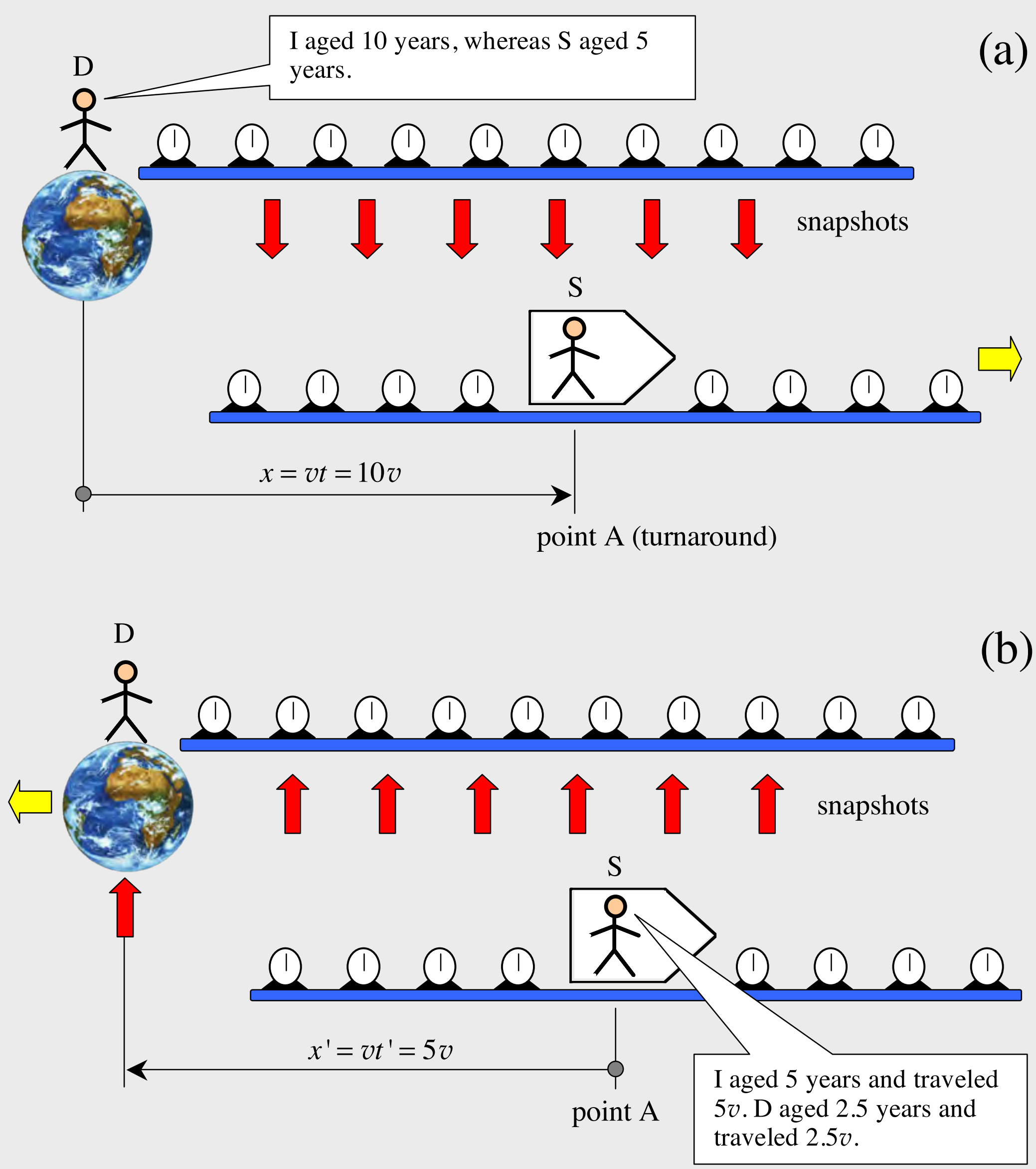


Figure 2. Measurements are carried out using lattices of clocks. (a) D is at rest. At t = 10 (x = $10v$), she records (using snapshots) distances and times in S's reference frame. Specifically, she takes snapshots of the clock readings in S's frame and finds t' = 5. She also finds that S's clocks are not synchronized with hers. (b) When S records t' = 5, her snapshots of D's lattice of clocks indicate that D's clock at point A reads 2.5.

Distances obey analogous relationships. S finds, using her lattice of clocks, that the earth-to-spaceship distance is larger than what D measures: $x' = vt' = v\gamma t = \gamma x$. Likewise, D, using her lattice of clocks, finds that the distance measured in S's frame is contracted: $x = \gamma x'$. Again, there is symmetry in the sense that D and S each find length contraction and time dilation ($\gamma = 2$) for space and time measurements carried out in the other person's inertial frame.

The (squared) interval Δs^2 is invariant with respect to boosts between inertial frames. When S gets to point A, she knows that $\Delta s^2 = (5c)^2$, because her clock reading is $t' = 5$ and she is at rest in her inertial frame. Likewise, she knows that from earth's inertial frame this interval is $(ct)^2 - x^2 = (ct)^2 - (vt)^2 = (ct)^2\gamma^{-2}$. Therefore, $t = 5\gamma = 10$ years. This argument also applies to the return trip, so S figures out that the time spent by D during the round trip is 20 years.

Ambiguity again arises when the system is examined from the other perspective. For example, with the earth moving relative to S's stationary reference frame, S uses her lattice of clocks to record times in both her frame and D's frame at exactly the turnaround point. In other words, S takes a snapshot of D's clock at the turnaround point. Not surprisingly, it is found that the system is completely symmetric in the sense that S discovers that 10 years has passed in her frame, whereas only 5 years has passed in D's frame. Indeed, D has aged only half as much as S at the turnaround point.

What distinguishes the twins – that is, breaks the symmetry of their respective measurements of time and length– is the turnaround at A, in which S switches from outbound to inbound. This holds the key to understanding what is going on.

Simultaneity

In D's frame, the ct and x axes are orthogonal and S's axes are slanted. In Fig. 3, the red dashed line parallel to the x' axis indicates simultaneity as measured with S's lattice of clocks. For example, according to S, at $t' = 5$ the time on earth is given by the intersection of the red dashed line and the ct axis. Keep in mind that the axes are scaled differently.

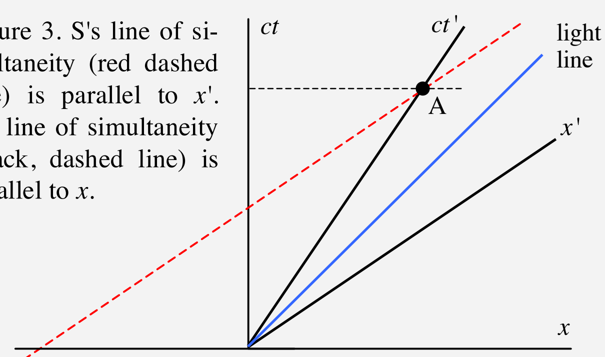


Figure 3. S's line of simultaneity (red dashed line) is parallel to x'. D's line of simultaneity (black, dashed line) is parallel to x.

There exists an infinity of red dashed lines of simultaneity parallel to the one shown, each corresponding to a different value of ct'. Lines of simultaneity are space-like. Events along them cannot be related causally, so observers using lattices of clocks need not agree on sequences of events.

It is impossible for two events to be judged as simultaneous from inertial frames in motion relative to one another, unless (trivially) the events have $x_1 = x_2$ and $t_1 = t_2$. In Fig. 3, an observer in the unprimed frame judges as simultaneous events lying on lines parallel to the x-axis. You might find it helpful at this point to revisit the discussion of simultaneity in Example 1.1. Figure 1.5 from this example is reproduced below as Fig. 4.

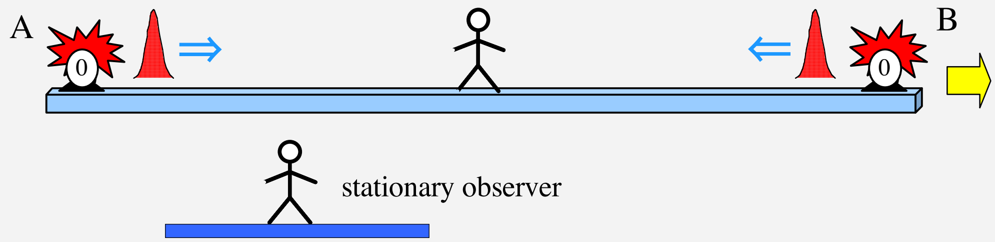


Figure 4. Synchronized clocks A and B emit brief pulses of light. An observer at rest mid-way on the platform receives the pulses simultaneously and concludes they were emitted simultaneously. The platform and its contents are in motion relative to a second observer who judges the bursts of light as not occurring simultaneously.

As mentioned earlier, despite the fact that S experiences acceleration during takeoff, turnaround, and descent, resolution of the twin paradox is possible without dealing explicitly with these brief periods. It is not as if acceleration is unimportant. After all, it accounts for the turnaround, and it can be introduced neatly into special relativity. Nonetheless, a calculation that takes explicit account of the acceleration needed for turnaround can be avoided. This is what is being done when it is assumed that turnaround is instantaneous. Another example of how to avoid such a calculation is that at point A, S could continue her outbound journey but transfer her clock to a passing spaceship traveling to earth at a constant speed of $0.866\,c$.[15]

From D's perspective, in which she is at rest in the unprimed frame, with S's spaceship moving in the $+x$ direction, D ages more rapidly than S. However, as long as they move apart from one another the situation is symmetrical. It does not matter which person is taken as being at rest, with the other person in motion. The only difference is the direction of the velocity. This symmetry includes takeoff and landing. For an age difference to be manifest this symmetry must be broken, for example, as happens with the round trip.

It is S who turns around and returns to earth. We could enlist a scenario in which S's spaceship is stationary and the earth first moves away from the spaceship and then turns around quickly and returns to it, but this is not done. As mentioned above, the choice in which S turns the spaceship around, while D remains stationary, breaks the symmetry. The time period over which the acceleration needed to turn the spaceship around takes place can be taken as irrelevant (arbitrarily small). What cannot be dismissed as irrelevant, however, is the fact that S can exchange her outbound lattice of clocks for an inbound one. These are the measuring devices she uses to keep track of events, including judgments of simultaneity or lack thereof.

Referring to Fig. 5, dashed red lines of simultaneity lie parallel to the respective x' axes for the outbound and inbound trips. S knows about time dilation, and she also knows how to figure out their ages and locations by using the invariance of the interval: Δs^2.

[15] This transfer can be carried out using photons to pass information when the spaceships are near one another.

For example, at point A, $(ct')^2 = (5c)^2 = (ct)^2 - (vt)^2 = (10c)^2 - (8.66c)^2$. In other words, when S gets to point A she can figure out, using a back-of-the-envelope calculation, that D will have aged 10 years. She also knows that D, using her own lattice of clocks, will find that S has aged 5 years. What S actually measures, however, using her lattice of clocks is that $t = 2.5$ at $x = 0$.

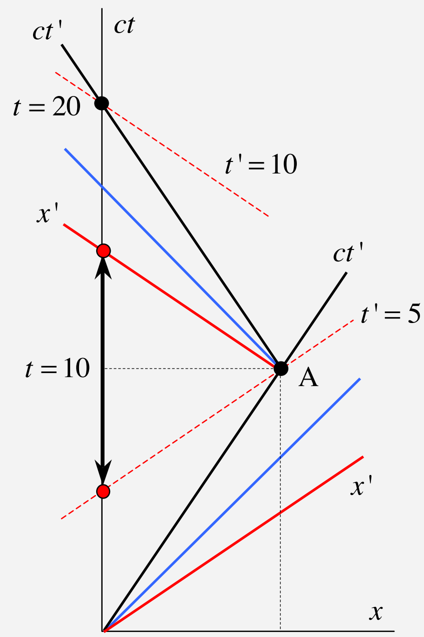


Figure 5. At the turnaround point A, S switches to a new lattice of clocks. Her lines of simultaneity (red) are parallel to the lower x' axis for the outbound trip and parallel to the upper x' axis for the inbound trip. The assumed instantaneous change of spaceship direction implies an instantaneous change of reference system (lattice of clocks). This causes S to overlook what happened in D's frame during the interval shown as the double-sided black arrow. This is a matter of judging simultaneity. Keep in mind that the ct and ct' axes are scaled differently, *e.g.*, at $t = 10$, $t' = 5$. Likewise, when S judges simultaneity along her $t' = 5$ line, she finds that $t = 2.5$, *i.e.*, $\Delta s^2 = (ct')^2 - (x')^2 = (5c)^2 - 0.866^2(5c)^2 = (5c)^2/4 = (2.5c)^2$. The lower red point is $t = 5$ when viewed by D, but it is 2.5 when S carries out a measurement of D's lattice of clocks.

The spaceship is turned around quickly at A, and in no time at all it is on a straight-line path traveling toward earth. Immediately following turnaround, the line of simultaneity for the return trip is the upper x' axis in Fig. 5. In other words, S's line of simultaneity undergoes a precipitous change at A.

Proper Time: Acceleration at A

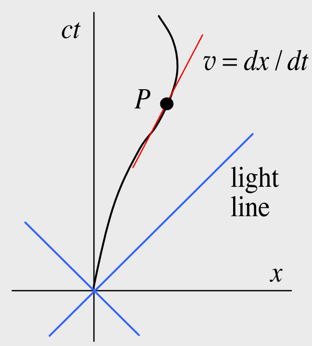
The issue of acceleration has been circumvented by assuming that it takes place over a small (infinitesimal) time interval. Suppose that we choose instead to deal explicitly with the acceleration that takes place in the vicinity of point A. This can be done as long as we enlist the perspective of an observer at rest who keeps track of what is going on in the accelerated frame. Lorentz transformation can then be used to see what this looks like from within the accelerated frame. This is analogous to what is done in Newtonian mechanics when Coriolis and centripetal terms are introduced into equations of motion that are valid in rotating (accelerated) frames.

The spaceship's acceleration in the turnaround region is equivalent to what it would experience in the presence of a strong gravitational field. Think in terms of 3 spatial coordinates rather than the x and x' used so far. Because of the acceleration, spacetime is no longer flat. The reference frame that accompanies the spaceship follows a geodesic on a curved surface. This results in a net reorientation of the reference frame after acceleration has ceased. This is like parallel transport on the surface of a sphere, specifically, a triad frame is parallel transported along a closed path constructed from geodesics. Such parallel transport is germane to geometric phase – quite a switch from space travel.

To proceed, consider an observer at rest ($x' = 0$) in a frame located at P in Fig. 6, and assume that this observer and her frame are moving at constant velocity relative to the unprimed frame. From Fig. 6, we see that, in fact, this observer is accelerated, as she does not have a straight world line. Locally, however, the line can be taken as straight in the sense that the correction to the velocity, dv, of the observer and her frame is small. The observer at P can make small corrections (of order dv) that enable v to change.

These changes can be calculated using nothing more elaborate than Newton's laws. In turn these results are transformed to the accelerated frame using Lorentz transformation. In this way, special relativity can be introduced to the world line of an accelerated observer. In our case, the accelerated observer is S as she guides the spaceship through its turnaround at point A.

Figure 6. The curved world line of an accelerated particle is approximated as a straight line at P. An observer in the unprimed frame notes the world line of the accelerated particle. Lorentz transformation enables the observer in the accelerated frame to record her proper time.



The math that goes with the above uses the invariance with respect to Lorentz transformation of the squared interval Δs^2. In the frame at P, which we take as primed, Δs^2 is given by $(cdt')^2$ because $x' = 0$. In the unprimed frame it is expressed as $c^2dt^2 - dx^2$. These facts yield

$$(cdt')^2 = c^2dt^2 - v(t)^2 dt^2 \ . \tag{1}$$

The fact that $v(t)$ can be obtained in the vicinity of P enables the proper time in the accelerated frame to be obtained by integration:

$$\int_{t_1'}^{t_2'} dt' = \Delta t' = \int_{t_1}^{t_2} dt \left(1 - v(t)^2/c^2\right)^{1/2} \ . \tag{2}$$

When S turns the spaceship around, her reference frame (lattice of clocks) undergoes reorientation until it is in place for the journey to earth at $v = 0.866\,c$. In the process, she would see time passing quickly on earth, were she inclined to check. However, she is busy guiding the spaceship through the turnaround maneuver. Referring to Fig. 5, by the time the turnaround has been completed her line of simultaneity has changed from the lower red dashed line to the upper solid red line. She has, in a sense, bypassed the ten-year period indicated by the black double-sided arrow in Fig. 5.

Optical Communication Between Frames

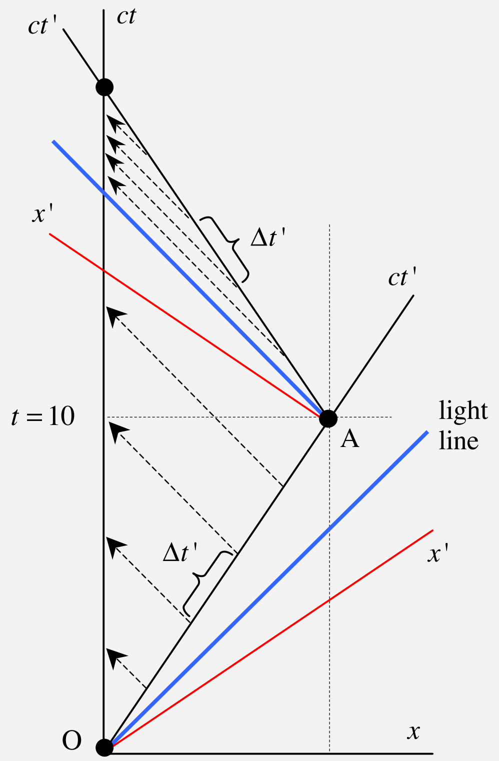
One way to understand the twin paradox that I find intuitive and easy to remember involves light that is transmitted between the unprimed and primed frames. The fact that frequencies are shifted through the Doppler effect makes it easy to distinguish the outbound and inbound parts of the trip. Two cases are considered: (1) monochromatic light emitted from S's laser arrives at D's detector; and (2) monochromatic light emitted from D's laser arrives at S's detector. We shall see how asymmetry and lack of simultaneity enter into the observations.

To begin, consider light sent from S to D, as indicated in Fig. 7 using dashed arrows. The diagram is from D's perspective; it is assumed that she and her reference frame are at rest. In a given time interval $\Delta t'$ in the moving (S's) frame, the number of light cycles is equal to $\nu_0 \Delta t'$, where ν_0 is the light frequency measured in the moving frame. These light cycles are received in the stationary frame in an interval Δt. For light emitted along the trip to point A, its frequency is lower in D's frame because the source is moving away from the observer. This was discussed in Example 2.4 entitled: *Doppler Effect*. The result is that Δt is given by

$$\Delta t = (1+\beta)\gamma \Delta t' . \tag{3}$$

In the non-relativistic limit ($\gamma = 1$, small β), this expression is recognized as following from the well-known expression for the Doppler shifted frequency for the case in which the source and observer are moving apart: $\nu \approx \nu_0 (1-\beta)$. On the inbound part of the trip, the frequency of the light received by D is higher by the Doppler shift, so Δt is given by: $(1-\beta)\gamma \Delta t'$.

Figure 7. From the starting point O to the turnaround point A, light emitted from the primed frame in a given time interval (say $\Delta t'$) arrives at the observer in the unprimed frame over a longer time interval $\Delta t = (1+\beta)\gamma\,\Delta t'$. Distances between dashed arrows correspond to one year in the primed frame. Keep in mind that the ct and ct' axes are scaled differently, *e.g.*, the horizontal line from $t = 10$ intercepts the ct' axis at $t' = 5$. After the turnaround, light sent from the primed frame to the unprimed frame arrives at the unprimed frame over a relatively short time interval. The observer in the unprimed frame can distinguish light arriving from the outbound versus inbound legs of the journey through the Doppler shifts of the received photons.



The time interval recorded by D for the radiation she detects that was emitted during the first half of S's trip is equal to the number of cycles multiplied by the duration of a cycle in D's frame. The latter is obtained using eqn (3) with Δt and $\Delta t'$ representing one cycle of light and summing $N/2$ cycles, where N is the total number of cycles sent by S over the entire round trip. Light sent by S on the inbound trip is compressed in time in D's frame (its frequency in D's frame is larger than ν_0), as indicated in Fig. 7. Summing the intervals associated with the outbound and inbound parts of the trip yields the total amount of time measured by D:

$$\underbrace{\tfrac{1}{2}N(1+\beta)\gamma\,\Delta t'_{\text{one cycle}}}_{\text{outbound}} + \underbrace{\tfrac{1}{2}N(1-\beta)\gamma\,\Delta t'_{\text{one cycle}}}_{\text{inbound}} = \gamma N \Delta t'_{\text{one cycle}} = \gamma\,\Delta t'_{\text{round trip}}\,. \qquad (4)$$

D has no problem figuring out what is happening. She records a total time of 20 years. From the Doppler shifts and knowledge of ν_0, she figures out that $\gamma = 2$. By counting the number of light cycles received over the 20-year period she confirms that S indeed sent out pulses over a 10-year period. So D is now 100% confident, though not happy, that she is the older twin. All that remains is to explain to S what is going on. To assist in this task, transmission in the opposite direction is examined.

S Receives Signals

Figure 8 illustrates the signals that S receives from D during the trip. Interestingly, even though S's trip requires 10 years, she receives light that was emitted over a span of 20 years.

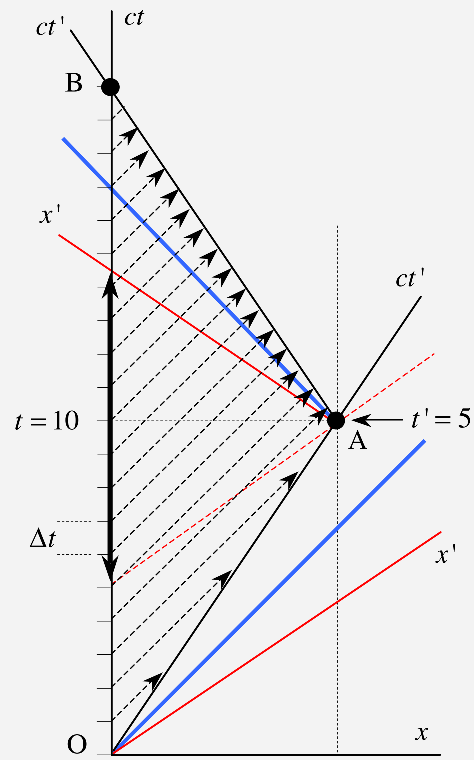


Figure 8. Starting at point O and following the *ct* axis to the end point B, light is emitted from the unprimed frame over a span of 20 years. A given time interval, say Δt, arrives at the primed frame over an interval $\Delta t'$ that is larger than Δt for the outbound trip and smaller than Δt for the inbound trip. Vertical distances between dashed arrows correspond to one year in the unprimed frame. Keep in mind that the *ct* and *ct'* axes are scaled differently. In the present case, the dashed horizontal line going to the right from $t = 10$ intercepts the *ct'* axis at $t' = 5$. Before the turnaround, S receives a relatively small percentage of the total number of photons sent by D. The majority of the photons arrive after the turnaround. When S adds up all the photons and takes into account their Doppler shifts, she finds that they were emitted over a 20-year period. In switching lattices of clocks, the interval shown with the double-sided black arrow (also in Fig. 3.6) is easy to overlook.

D and S now agree on what happened. As mentioned earlier, the key to the resolution of the ambiguity resides in a careful treatment of the turnaround. If one only considers the intervals from O to just before A, and from just after A to B, then D and S each age by 10 years. During the period of intense acceleration near A, S's clock slows considerably, so much so that 10 years pass on earth during this time. She figured this out during the return trip by analyzing the Doppler shifted photon flux that she received from D.

This effect has been confirmed experimentally. Two precision (cesium) clocks were synchronized, one was sent around the earth on a great circle path, and their times were compared when they were again in the same location.

Proper Time Tells All

An understanding of proper time enables the twin paradox to be settled without convoluted arguments or subtle math. Let us carry this out from D's perspective. Her proper time is 20 years. From her frame we now compute the proper time for S's route. This is equal to $2(100-75)^{1/2} = 10$ years. That is all there is to it.

The twin paradox is, for many people, made more amusing if a mother and her daughter are used instead of twins. The mother travels on the space ship, while her daughter stays behind on earth. The daughter is skeptical of special relativity, and is also less adventurous than her mother. When they meet at the end of the spaceship's round trip, the mother is younger that her daughter. This raises questions that range from causality to religion.

3.3. Velocity and Momentum

In this subsection the Lorentz covariant four-velocity U^{ν} is defined, and from there the four-momentum for massive particles P^{ν} falls neatly into place. Following this, an alternate (complementary) route to P^{ν} is presented as Example 3.3. We then move on to define the four-acceleration B^{ν} and the four-force F^{ν}. The latter is used to obtain two well-known expressions for the relativistic energy of a massive particle: $E = \gamma mc^2$ and $E^2 = p^2c^2 + m^2c^4$. These are easily generalized to include the photon through the enlistment of a heuristic argument.

We shall borrow some nomenclature from Sections 4 and 5 in the material that follows immediately below. The alternative would be to stick with a nomenclature that would suffice, but would be abandoned as soon as we got to Sections 4 and 5, to say nothing of the fact that the mathematics would be less efficient. If the material below seems vague (more so than usual), bear with it for the time being. On the other hand, it would be smart to read Sections 4 and 5, or at least the most relevant parts, before proceeding.

The four-velocity is defined as

$$U^{\nu} \equiv c\frac{dx^{\nu}}{ds} = \frac{dx^{\nu}}{d\tau} = \gamma\frac{dx^{\nu}}{dt}. \tag{3.6}$$

where $d\tau = ds/c$ and $dt = \gamma d\tau$ have been used. The Greek generic superscript ν denotes what are referred to as the contravariant components, with $\nu = 0$ corresponding to ct and $\nu = 1, 2, 3$ corresponding to the usual x, y, z.

It is necessary to differentiate x^{ν} with respect to proper time τ in order to have a sensibly defined relativistic velocity, namely, a Lorentz covariant four-vector. Lack of simultaneity would make a shambles of any attempt to use dx^{ν}/dt as the relativistic velocity. The reason is that time intervals, including differential elements of time, are frame dependent, unlike in the non-relativistic regime. Because the proper time τ is a Lorentz scalar, U^{ν} transforms in the same manner as the four-vector x^{ν}. Thus, U^{ν} is identified as being a four-vector. Keep in mind that the frame velocity $\vec{v}$ that was introduced in Section 1 is obviously not a four-vector. We will find that it is useful to express four-acceleration and four-force in terms of 3D quantities.

The proper time is the time measured in the rest frame of the particle or object. In this frame, there is, by definition, no three-velocity. Consequently, γ is equal to 1, and $U^{\nu} = U^0 = dct/d\tau = c$. In general, U^0 is equal to γc, but the fact that $\gamma = 1$ in the case of the particles' rest frame means that $U^0 = c$. Of course, it is never possible for U^0 to vanish, as time cannot stand still. Thus, in the particle's rest frame we immediately reveal a useful scalar: $U^{\nu}U_{\nu} = c^2$. It is a simple matter to transform the four-velocity from its rest frame to any inertial frame by using a Lorentz boost.

The components U^ν transform according to

$$U^{\nu'} = \Lambda^{\nu'}{}_{\varepsilon} U^{\varepsilon}, \tag{3.7}$$

where $\Lambda^{\alpha'}{}_{\varepsilon}$ is the boost. So far we have focused on the boost part of the Lorentz group, but rotation will enter soon (Section 6). As mentioned earlier, the discrete symmetries of parity and time reversal enter in the quantum mechanical version discussed in Chapter 4.

Recall from Section 1 the column vector expression for a spacetime coordinate four-vector in terms of ct, x, y, and z:

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}. \tag{3.8}$$

If we are to use this vector to compute the squared interval by contracting it with a row vector, the row vector must contain minus signs for the spatial dimensions in order to deliver the correct form for the squared interval. This is achieved by multiplying the above column vector from the left by $[ct, -x, -y, -z]$. The vector with the minus signs (which can be written as either a column or row vector) is the covariant version. Its components are written with subscripts, U_ν. Note from the box that the scalar product $U^\nu U_\nu$ (implied summation) is equal to c^2, as we saw earlier.

$$\begin{aligned} U^\nu U_\nu &= \frac{c^2 dt^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2}{d\tau^2} \\ &= \frac{c^2 dt^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2}{ds^2 / c^2} \\ &= c^2 \end{aligned}$$

The four-momentum scalar product for massive objects is $P^\nu P_\nu = m^2 U^\nu U_\nu = (mc)^2$.

We are now in a position to write expressions for the four-momentum P^ν. It is obtained from the four-velocity by multiplying U^ν by mass m. In some books (usually older ones), a subscript zero is appended to m, and the resulting m_0 is referred to as the rest mass. This is not a good practice. The mass is always the same; it does not depend on velocity. Though many of the equations that we encounter are of such a nature that mass might appear to increase with velocity, this is just mathematical appearance. Mass is a scalar that transforms intact between inertial frames, so we will use fixed m. The only time we need to be careful about m not remaining fixed is when particles are created and annihilated, which happens at quite high energy.

The introduction of mass in this manner into the four-momentum carries with it an implicit assumption that we are dealing only with massive objects. This excludes photons. However, we will see that the photon can be recovered in the limit $m \to 0$ by enlisting a straightforward, albeit heuristic, argument. Referring to eqn (3.6), the space components of the four-momentum are

$$P^i = mU^i = \gamma m \frac{dx^i}{dt}, \tag{3.9}$$

where $i = 1, 2, 3$. For the zero-component, differentiating ct with respect to time yields c, so P^0 is given by

$$P^0 = mU^0 = \gamma m \frac{d(ct)}{dt} = \gamma mc\,. \tag{3.10}$$

The momenta P^1, P^2, and P^3 have a fairly obvious role, whereas P^0 is more interesting. To see why it is more interesting, consider the kinetic energy of a particle of mass m. Its kinetic energy is given by the integral of the space part of the relativistic four-force along the direction of motion. The steps leading to an expression for the kinetic energy are carried out in a box a few pages from here. From eqn (iv) in the box (*vide infra*) it follows that when the velocity is zero, the kinetic energy is zero. The expression for the total energy is seen to contain the term mc^2. This is referred to as the rest energy, and the total energy is given by

$$E = \gamma mc^2\,. \tag{3.11}$$

Thus, referring to eqn (3.10), we see that $P^0 = E / c$.

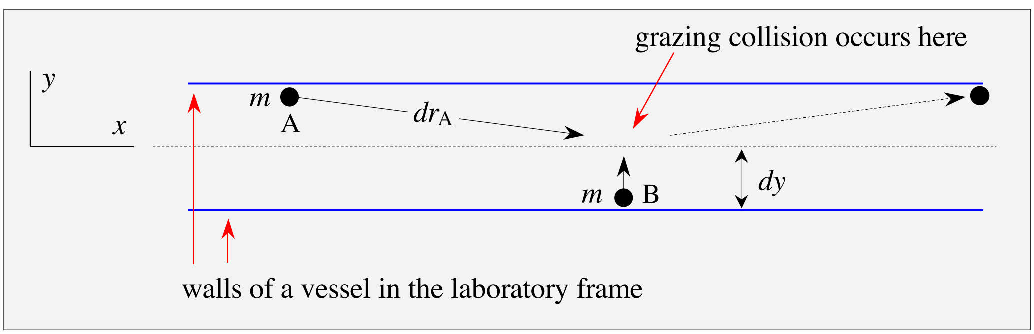
Example 3.2. Relativistic Momentum via a Simple Model

Before deriving the expression for relativistic energy, let us complement the material presented in the previous subsection. The following derivation of relativistic momentum is adapted from one in Taylor and Wheeler [5].

To start, consider something as elementary as a particle's displacement due to its momentum. The particle's displacement in a given direction is always due to its momentum in that direction. If this were not the case – suppose momentum at an angle θ to the direction of displacement is responsible for the displacement – there would be a continuum of momenta (along the ϕ coordinate at polar angle θ) that did the same thing. This is impossible in isotropic space. Thus, momentum and velocity point in the same direction. This turns out to be relevant in the discussion that follows.

A useful construct for our purposes is one in which a grazing collision takes place between two particles each of mass *m*, as indicated in the sketch below. One of the particles moves slowly, both before and after the collision, whereas the other particle moves rapidly, both before and after the collision. The strategy is this. The proper time of the slow particle, for all practical purposes, is equal to the particle's laboratory time, that is, the particle's motion lies safely in the Newtonian limit. This is a no-brainer. For example, $v = 10^{-3}c$ gives $\gamma \approx 1.0000005$.

Momentum conservation is used to relate the slow particle's momentum, which is exclusively in the *y*-direction, to the *y*-component of the fast particle's momentum. The invariance of proper time is used to link different inertial frames. The fast particle can be referenced to (1) the lab frame, (2) a frame whose velocity is equal to the particle's laboratory *x*-component, or (3) the particle's rest frame. We then return to the fact that displacement in a given direction is due to momentum in that same direction in order to enlist similar triangles and complete the derivation.



Referring to the above sketch, the collision is arranged such that the *y*-component of the system's overall momentum in the lab frame is zero. Thus, the *y*-component of particle A's momentum has the same magnitude as the *y*-component of particle B's momentum. Moreover, these magnitudes are conserved in the grazing collision. Particle B moves upward with momentum $p_{y,\mathrm{B}}\,\hat{y}$, undergoes a grazing collision with particle A, and then moves downward with momentum $-\,p_{y,\mathrm{B}}\,\hat{y}$. Think of billiards. Again, given that particle B travels slowly, its laboratory time can be taken as equal to its proper time.

Applying conservation of momentum to the y-direction, the y-component of particle A's momentum has magnitude $p_{y,\mathrm{A}} = p_{y,\mathrm{B}} = mdy/d\tau$, where $d\tau$ is the time it takes the particle to traverse the distance dr_{A}, as measured in the particle's rest frame. You may wish to think of particle A's motion in a reference frame that is moving toward the right with the same speed as the x-component of particle A's lab velocity. In this frame, particle A moves downward slowly with proper time τ that can be taken as equal to the particle's transit time measured in the x-directed frame. Thus, the proper times of particles A and B are the same, which is fairly obvious. This proper time is the same in any inertial frame, so we then ask if the quantity $mdr_{\mathrm{A}}/d\tau$ is in fact particle A's momentum.

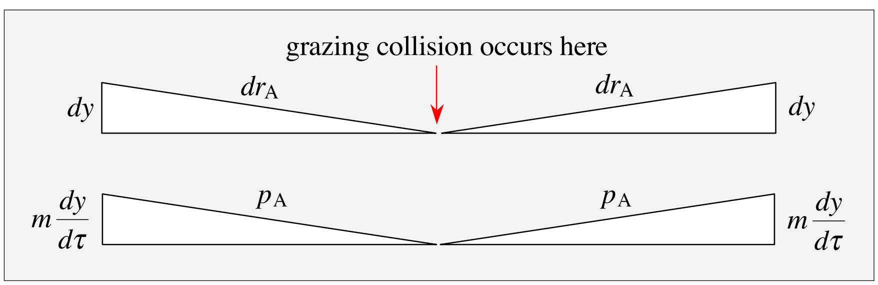
The triangles in the sketch below illustrate displacement and momentum of particle A in the laboratory reference frame. Because the displacements are in the directions of the momenta, the triangles are similar. Consequently, $p_A = m\,dr_A/d\tau$, and we can write

$$\vec{p}_A = m\frac{d\vec{r}_A}{d\tau}\,, \tag{1}$$

with momentum components:

$$p^i = m\frac{dx^i}{d\tau}\,. \tag{2}$$

$$p^0 = m\frac{d(ct)}{d\tau} = \gamma mc = E/c\,. \tag{3}$$

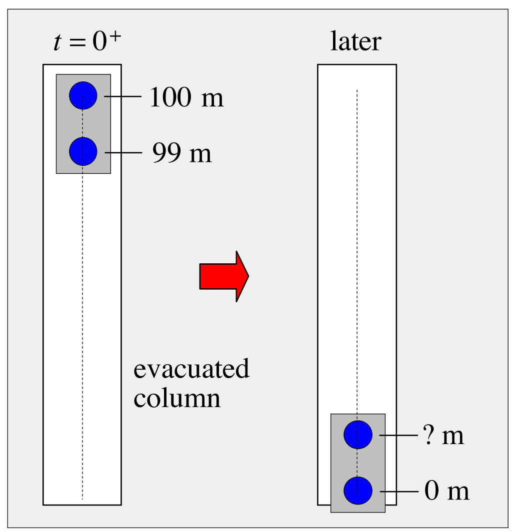


Now onward to acceleration, force, and energy in the relativistic regime:

3.4. Acceleration and Force

The theory of general relativity (gravitation) is impressive to say the least – one of the best theories ever invented. Though it applies to everything, truth be told, it deals almost exclusively with vastly larger length scales than the microscopic ones where relativistic quantum field theory and quantum mechanics play out. In sharp contrast to Newton's gravitational law, in Einstein's theory of general relativity, gravitational force is nowhere to be seen. Effects that we usually think of as due to gravitational force are accounted for by the curvature of spacetime.

General relativity is fascinating to the point of seduction. Nonetheless, in subsequent chapters we shall stick to electrodynamics and relativistic quantum theory, and these are rooted firmly in special relativity. There are cases in high-resolution spectroscopy where general relativity needs to be taken into account in dealing with atoms, but for the most part we can safely ignore gravitation at atomic and molecular length scales. Of course, we already know that at these length scales special relativity works remarkably well here on earth, where $g \approx 9.8$ m s^{-2}.



A number of examples are given in Taylor and Wheeler [5] that illustrate the degree to which gravity violates the assumptions upon which special relativity is premised. Usually this violation is because the chosen inertial frames turn out to be not quite inertial. One such example that makes the point nicely is the following: Two metal spheres inside an imaginary gray box that serves as their local inertial frame are dropped at the same time ($t = 0^+$) from a completely evacuated vertical column (sketch on the right). One is released at a height of 100 meters, the other at a height of 99 meters. They move toward earth along the same straight line.

How far apart are the spheres when the lower sphere touches the ground? Of course, were they in a truly inertial frame, they would be 1 meter apart. However, because of the r^{-2} dependence of the gravitational force, the finite distance between them translates into each of them experiencing a slightly different gravitational force at a given time. Go ahead and write a few equations and plug in some numbers. You will find that their initial separation of 1 meter is affected, but only by a tiny amount. The bottom line is that – though the box is not a truly inertial frame – it behaves like one to a high degree of approximation.

Four-Acceleration

The four-acceleration, B^ν, is defined as the derivative with respect to proper time of the four-velocity:

$$B^\nu \equiv \frac{dU^\nu}{d\tau} . \tag{3.12}$$

As such, it transforms covariantly as a four-vector. Our choice of the symbol B^ν for four-acceleration (rather than A^ν) is in deference to A^ν labeling the electromagnetic gauge field that occupies center stage in Chapter 2. The contraction on Minkowski space, $B^\nu B_\nu$, is an invariant. It has the same value in all inertial frames.

There are many fascinating and unanticipated features of acceleration at high speed. In the present section, we shall see how B^ν relates to the familiar non-relativistic three-velocity, $\dot{\vec{r}}$, and three-acceleration, $\ddot{\vec{r}}$, and a couple of examples will be discussed.

Let us begin by differentiating eqn (3.6) with respect to the proper time τ. Careful bookkeeping is needed because a particle's acceleration is not necessarily in the same direction as its velocity, certainly not in electrodynamics. For a particle at rest at the origin of a moving (primed) frame, its four-velocity relative to a stationary (unprimed) frame is $U^\nu = (\gamma c, \gamma \vec{v})$.

Differentiation of U^ν with respect to τ acts on both γ and $\vec{v}$. The former yields

$$\begin{aligned} \frac{d\gamma}{d\tau} &= -\tfrac{1}{2}\underbrace{\left(1-\frac{v^2}{c^2}\right)^{-3/2}}_{\gamma^3}\left(-\frac{2\gamma}{c^2}\vec{v}\cdot\underbrace{\frac{d\vec{v}}{dt}}_{\vec{a}}\right) \\ &= \gamma^4\frac{(\vec{v}\cdot\vec{a})}{c^2} . \end{aligned} \tag{3.13}$$

As promised, $\vec{v}$ and $\vec{a}$ are being used as non-relativistic ingredients of relativistic four-acceleration B^ν. Equation (3.13) is now used with $dU^\nu / d\tau$ to obtain the time and space parts of the four-acceleration, B^0 and $\vec{B}$, respectively:

$$B^0 = \frac{dU^0}{d\tau} = c\frac{d\gamma}{d\tau} = \gamma^4\frac{(\vec{v}\cdot\vec{a})}{c} \tag{3.14}$$

and

$$\vec{B} = \frac{d(\gamma\vec{v})}{d\tau} = \gamma^4\frac{(\vec{v}\cdot\vec{a})}{c^2}\vec{v} + \gamma^2\vec{a} . \tag{3.15}$$

Taking the inner product of B^ν with B_ν on Minkowski space yields

$$B^\nu B_\nu \;=\; \gamma^8 \frac{(\vec{v}\cdot\vec{a})^2}{c^2} - \left(\gamma^4 a^2 + 2\gamma^6 \frac{(\vec{v}\cdot\vec{a})^2}{c^2} + \gamma^8 \frac{(\vec{v}\cdot a)^2}{c^4} v^2 \right) \tag{3.16}$$

$$= -\gamma^4 a^2 + \gamma^8 \frac{(\vec{v}\cdot\vec{a})^2}{c^2}\left(1 - \frac{v^2}{c^2}\right) - 2\gamma^6 \frac{(\vec{v}\cdot\vec{a})^2}{c^2} \tag{3.17}$$

$$= -\gamma^4 a^2 - \gamma^6 \frac{(\vec{v}\cdot\vec{a})}{c^2}\,. \tag{3.18}$$

Earlier it was noted that velocity and acceleration (force) in general are not collinear. In fact, often they are at right angles to one another. For example, in the Lorentz force equation of electrodynamics: $\vec{F} = q(\vec{E} + (\vec{u}\,/\,c)\times\vec{B})$, the term $q(\vec{u}\,/\,c)\times\vec{B}$ is a force that is perpendicular to the velocity $\vec{u}$. To make such geometrical features conspicuous, use is made of a vector identity that you will find on the first page of the handout entitled Miscellaneous Math, specifically, $(\boldsymbol{a}\times\boldsymbol{b})\cdot(\boldsymbol{c}\times\boldsymbol{d}) \;=\; (\boldsymbol{a}\cdot\boldsymbol{c})(\boldsymbol{b}\cdot\boldsymbol{d}) - (\boldsymbol{a}\cdot\boldsymbol{d})(\boldsymbol{b}\cdot\boldsymbol{c})$. This enables eqn (3.18) to be written

$$B^\nu B_\nu \;=\; -\gamma^4 a^2 - \frac{\gamma^6}{c^2}\left(v^2 a^2 - (\vec{v}\times\vec{a})^2\right)$$

$$= -\gamma^4 a^2 \left(1 + \gamma^2 \frac{v^2}{c^2}\right) + \frac{\gamma^6}{c^2}(\vec{v}\times\vec{a})^2$$

$$= -\gamma^6 \left(a^2 - \frac{(\vec{v}\times\vec{a})^2}{c^2} \right) \tag{3.19}$$

In the non-relativistic limit ($\vec{v}\,/\,c \to 0$), $B^\nu B_\nu$ becomes $-a^2$, as expected. Also, we see that $B^\nu B_\nu$ is equal to $-\gamma^6 a^2$ when $\vec{v}$ and $\vec{a}$ are parallel to one another, whereas $B^\nu B_\nu$ is equal to $-\gamma^4 a^2$ when $\vec{v}$ and $\vec{a}$ are perpendicular to one another.

The four-vector acceleration B^ν has a number of properties that differ qualitatively from those of U^ν and P^ν. An obvious difference is that $B^\nu B_\nu$ is negative, whereas $U^\nu U_\nu$ and $P^\nu P_\nu$ are positive. In fact, B^ν is orthogonal to U^ν, which can be verified with a simple calculation using $U^\nu \;=\; (\gamma c, \gamma\vec{v})$ and eqns (3.14) and (3.15). Thus, B^ν is a spacelike four-vector. That is, it lies outside the light cone. One of the more interesting aspects of the four-acceleration is that it achieves the non-relativistic limit through velocity, not acceleration, as seen in eqns (3.14) and (3.15). As $\vec{v}\to 0$, B^0 vanishes and $\vec{B}$ becomes $\vec{a}$. Thus, even quite large accelerations can be dealt with non-relativistically as long as $v\,/\,c$ is negligible.

Exercise: Show that $B^\nu U_\nu = 0$. Explain why the fact that $B^\nu B_\nu$ is negative tells us immediately that B^ν is a spacelike vector.

The four-force F^ν behaves very much like B^ν, as it is nothing more than B^ν multiplied by mass, *m*. Therefore, it has the property that $F^\nu F_\nu$ is negative, and the four-force is spacelike. We have seen that $B^\nu U_\nu = 0$, but we need to be more careful with $F^\nu U_\nu$. The reason is that $F^\nu = mB^\nu$ contains the particle mass *m*. The mass of a high-energy system can change due to the creation and annihilation of particles. For example, a high-energy electron that suddenly experiences a strong potential can create electron-positron pairs. Particle creation and annihilation are dealt with using relativistic quantum field theory. Though not part of the classical theory discussed in this chapter, it is something to keep in mind. Aside from processes like this, we are free to use $F^\nu U_\nu = 0$.

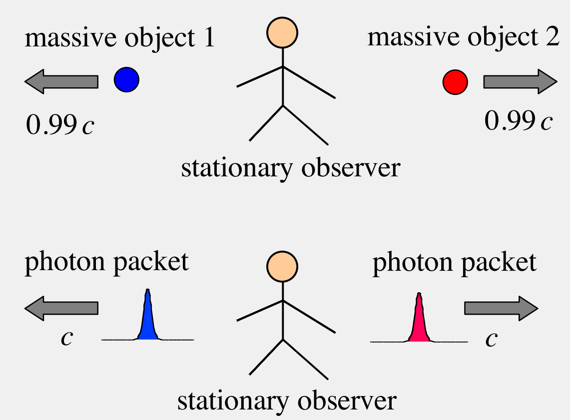
So far we have focused on a particle that experiences force. Let us now examine a particle that is in motion and experiences acceleration with respect to two inertial frames, one of which we shall take as stationary. This leads to the interesting case of a particle that experiences constant acceleration, which is the subject of Example 3.4. Before continuing, you may wish to review Section 2.2. *Relating Velocities Between Frames*.

In the material presented above, four-vectors and their properties were expressed in terms of non-relativistic three-vectors. In the next two subsections we shall work directly with three-vector quantities, as this is used in the graphical interpretation given at the end.

3.5. Mutual Velocity

Let us begin by noting the difference between two seemingly similar situations. In one, an observer measures the velocities of two particles using her lattice of clocks. She then compares these velocities to see how rapidly the distance between the particles increases as a function of time. In the other situation, she measures the velocity, relative to her, of an object or a point in space. The second situation applies to nearly all of the cases that we have examined so far.

Referring to the sketch on the right, suppose the observer stands still and records the motions of two massive objects. She finds that massive object 1 is moving away from her in the $-x$ direction at a speed close to c. In addition, she finds that massive object 2 is moving away from her in the $+x$ direction at a speed close to c. Thus, the observer finds that the distance between these objects is changing at a speed close to $2c$. This is an example of what is called the *mutual velocity* of two objects. How it acquired this moniker is beyond me. The mutual velocity requires the measurement of more than one object. Note that the "objects" need not be massive. For example, the above description and sketch apply equally well to a pair of short duration photon packets traveling in opposite directions, as indicated in the sketch.



The observer finds that the time rate of change of the distance between the two massive objects is $1.98c$ ($2c$ for the photon packets).

The mutual velocity is obviously different than the *relative velocity* of two objects or the velocity of an object relative to a point in space. These can never exceed c. If an observer sitting on object 1 measures the speed of object 2, it will always be less than c. This is a measurement of the relative velocity. Now suppose that object 2 moves at a speed of almost c in the $+x$ direction with respect to a moving reference frame. Furthermore, assume that, in turn, this reference frame moves at a speed of almost c in the $+x$ direction with respect to a stationary observer. This observer never measures that the object is moving at a speed of c or higher. Recall the velocity addition formulas derived earlier. Again, this is a measurement of the relative velocity.[16]

The concept of mutual velocity is easy to accept once it is appreciated that it is entirely

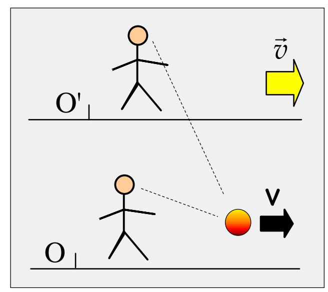
[16] On the other hand, phase velocities easily exceed c. Think of the intersection point of a pair of scissors, wave fronts incident on a beach at near-normal incidence, or the beam from a laser pointer scanned rapidly across a wall. In such cases, no object or electromagnetic energy is transported at a speed in excess of c.

consistent with special relativity. It arose earlier when Arman asked about it in the context of the addition of velocities. Greater attention is required with acceleration. To get a glimpse of how this works, let us see how accelerations are viewed from two different reference frames.

Object Moves Relative to Two Frames

To begin, we write expressions for velocities relative to two inertial frames, say, the usual unprimed and primed ones:

$$\vec{u} = \left(\frac{dx^1}{dt}, \frac{dx^2}{dt}, \frac{dx^3}{dt} \right) \qquad \vec{u}\,' = \left(\frac{dx^{1'}}{dt'}, \frac{dx^{2'}}{dt'}, \frac{dx^{3'}}{dt'} \right) \tag{3.20}$$



Lorentz transformation is now used to express the primed velocity components in terms of the unprimed velocity components: $dt' = \gamma(dt - (v/c^2)dx^1)$; $dx^{1'} = \gamma(dx^1 - vdt)$; $dx^{2'} = dx^2$; and $dx^{3'} = dx^3$. As usual, the velocity of the primed frame relative to the unprimed frame, $\vec{v}$, is assumed to be in the $+x^1$ direction. A bit of minor algebra yields[17]

$$\vec{u}\,' = D^{-1}\left(u^1 - v,\, \gamma^{-1} u^2,\, \gamma^{-1} u^3\right), \tag{3.21}$$

---

[17] The Lorentz transformation of coordinate intervals, in matrix form, is

$$\begin{bmatrix} dx^{0'} \\ dx^{1'} \\ dx^{2'} \\ dx^{3'} \end{bmatrix} = \begin{bmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} dx^0 \\ dx^1 \\ dx^2 \\ dx^3 \end{bmatrix} = \begin{bmatrix} \gamma dx^0 - \beta\gamma dx^1 \\ -\beta\gamma dx^0 + \gamma dx^1 \\ dx^2 \\ dx^3 \end{bmatrix}.$$

Taking the ratios indicated in eqn (3.20) yields

$$u^{1'} = \frac{dx^{1'}}{dt'} = \frac{\gamma(dx^1 - vdt)}{\gamma(dt - \frac{vdx^1}{c^2})} = \frac{u^1 - v}{1 - \frac{vu^1}{c^2}} = D^{-1}(u^1 - v)$$

where $D = 1 - vu^1/c^2$

$$u^{2'} = \frac{dx^{2'}}{dt'} = \frac{dx^2}{\gamma\left(dt - \frac{vdx^1}{c^2}\right)} = \frac{u^2}{\gamma\left(1 - \frac{vu^1}{c^2}\right)} = D^{-1}\gamma^{-1}u^2$$

$$u^{3'} = \frac{dx^{3'}}{dt'} = \frac{dx^3}{\gamma\left(dt - \frac{vdx^1}{c^2}\right)} = \frac{u^3}{\gamma\left(1 - \frac{vu^1}{c^2}\right)} = D^{-1}\gamma^{-1}u^3.$$

This is eqn (3.21).

where $D = 1 - vu^1 / c^2$.

This generalizes a result obtained in Section 2.2. *Relating Velocities Between Frames*. There, it was assumed that a point moves in the $+x^1$ direction with speed v relative to a stationary frame, and this point is observed from a moving frame whose velocity relative to the stationary frame is $v\hat{x}^1$. Figure 1.18 from that section is repeated here.

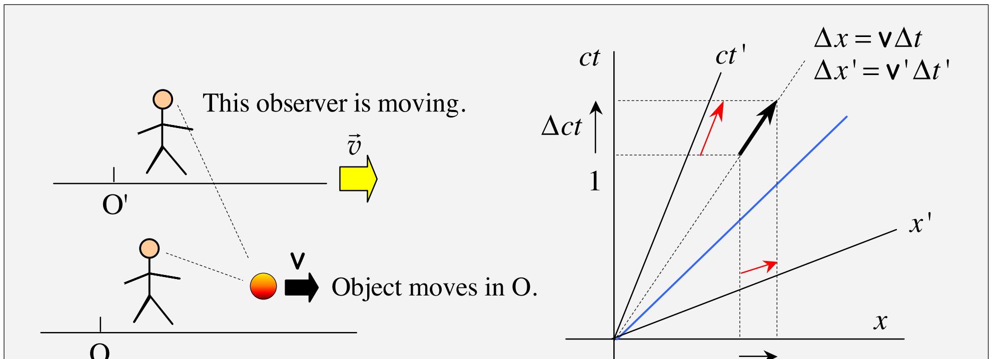


Figure 1.18 from Section 2.2

If we compare eqn (1.23) in this earlier section:

$$\mathsf{v}' = \frac{\mathsf{v} - v}{1 - v\mathsf{v} / c^2}, \tag{1.23}$$

to eqn (3.21), we find that these equations are identical for the case $\vec{u}' = D^{-1}(u^1 - v)\hat{x}^1$ ($u^{2'} = u^{3'} = 0$). This material was also covered in Section 2.3, Matrix Representation: Velocity Addition. The velocity addition formula derived there, eqn (2.15), is repeated below. It is the same as eqn (1.23) when one takes into account the different direction of the velocity $\vec{v}$. In other words, the v in eqn (1.23) must be replaced by $-v$ to be compatible with eqn (2.15).

$$v'' = \frac{v + v'}{1 + vv'/c^2}. \tag{2.15}$$

The bottom line is that we have reviewed some material, and we have derived a couple of expressions that will now be used to examine acceleration as seen from different frames and by an observer undergoing proper acceleration.

3.6. Constant Acceleration and Hyperbolic Motion

The acceleration measured from the primed frame: $d\vec{u}'/dt'$, is obtained by dividing the differential velocity element $d\vec{u}'$ by dt'. The $du^{1'}$ component of the former is obtained using eqn (3.21):

$$
\begin{aligned}
du^{1'} &= D^{-2}\left(Ddu^1 - (u^1 - v)dD\right) \\
&= D^{-2}\left(Ddu^1 + (u^1 - v)\frac{v\,du^1}{c^2}\right),
\end{aligned} \tag{3.22}
$$

where $D = 1 - u^1 v / c^2$. Dividing this equation by dt' yields the $x^{1'}$ component of the acceleration seen from the primed frame:

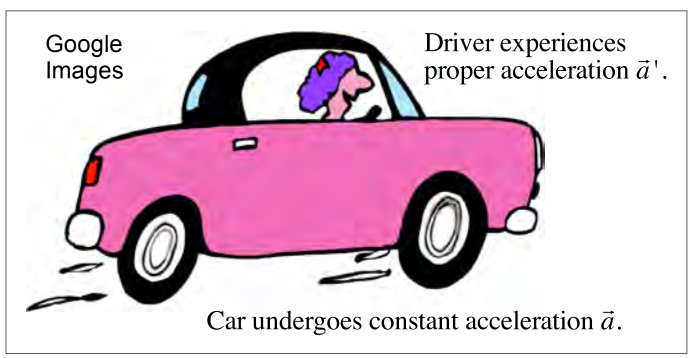
$$
a^{1'} = \frac{du^{1'}}{dt'} = \frac{1}{\gamma D^2}\frac{Ddu^1 + (u^1 - v)\dfrac{vdu^1}{c^2}}{dt - \dfrac{v\,dx^1}{c^2}},
$$

and with minor algebra this becomes

$$
a^{1'} = \frac{a^1}{\gamma D^3}\left(D + (u^1 - v)\frac{v}{c^2}\right) = \frac{a^1}{\gamma^3 D^3}. \tag{3.23}
$$

Exercise: Referring to eqns (3.21) – (3.23), derive the corresponding expressions for $a^{2'}$ and $a^{3'}$.

Let us now examine the special case of a particle (or anything else for that matter) that is at rest at the origin of the moving frame. In this case, $u^1 = v$, and $u^2 = u^3 = 0$. The acceleration experienced by the particle, $a^{1'}$, is its proper acceleration. For example this is the acceleration that would be experienced by a person in a car that is undergoing acceleration, as illustrated in the cartoon. For the case $u^1 = v$, D is equal to γ^{-2}, and eqn (3.23) becomes



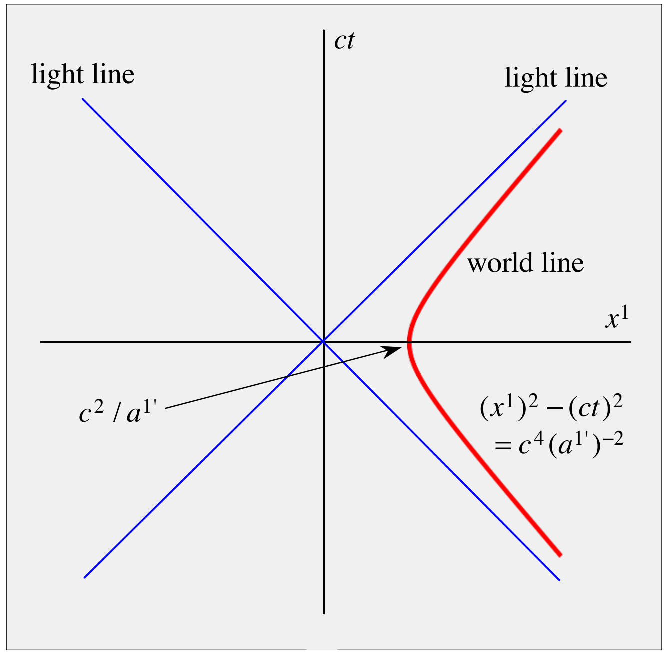
$$
a^{1'} = \gamma^3 a^1 = \gamma^3 \frac{du}{dt} = \frac{d(\gamma u)}{dt} \tag{3.24}
$$

Exercise: Verify eqn (3.24).

When $a^{1'}$ is constant, integration of eqn (3.24) does not present a conceptual challenge, though patience is essential.[18] This yields

$$(x^1)^2 - (ct)^2 = c^4 (a^{1'})^{-2} . \tag{3.25}$$

This is the equation of a hyperbola (see sketch on the right). We recognize the left hand side of eqn (3.25) as the negative of the squared interval $(ct)^2 - (x^1)^2$, with the right hand side having the same value in every inertial frame. Go ahead and add one or more sets of primed axes to the sketch. The particle's world line is the red hyperbola. Note that the distance of closest approach to the apex is $c^2 / a^{1'}$.



Starting from the lower right, the particle has a large negative velocity that is close to c. As it approaches the point $x^1 = c^2 / a^{1'}$, it is slowed and reaches a turning point at $t = 0$, where its velocity is zero. It then proceeds to increase in the x^1 direction, and at long times its speed gets close to c.

---

[18] The math is tedious but not difficult. Start by integrating $a^{1'}dt = d(\gamma u)$, with $u = 0$ at $t = 0$, to obtain $a^{1'}t = \gamma u$. Then square and solve for u, using $\gamma^{-2} = 1 - u^2 / c^2$:

$$(a^{1'}t)^2 = \frac{u^2}{1 - u^2 / c^2} \Rightarrow u^2 \left(1 + (a^{1'}t / c)^2\right) = (a^{1'}t)^2 \Rightarrow u = a^{1'}t \left(1 + (a^{1'}t / c)^2\right)^{-1/2} .$$

Now integrate $u = dx^1 / dt$: $x^1 - x(0) = \int dt \, \frac{a^{1'}t}{\sqrt{1 + (a^{1'}t / c)^2}} = \frac{c^2}{a^{1'}} \int d(a^{1'}t / c) \, \frac{a^{1'}t / c}{\sqrt{1 + (a^{1'}t / c)^2}}$,

where $x(0) = c^2 / a^{1'}t$.

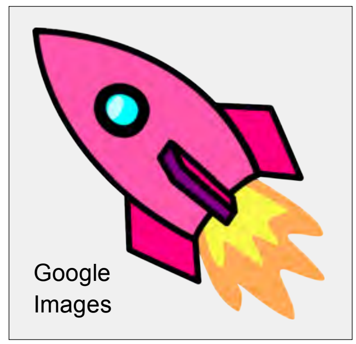
With $y = a^{1'}t / c$, this reads: $x^1 - x(0) = \frac{c^2}{a^{1'}} \int dy \, \frac{y}{\sqrt{1 + y^2}}$. Now use $1 + y^2 = \beta$ to obtain

$$x^1 - x(0) = \frac{c^2}{2a^{1'}} \int d\beta \, \beta^{-1/2} = \frac{c^2}{a^{1'}} \beta^{1/2} = \frac{c^2}{a^{1'}} \left(\sqrt{1 + (a^{1'}t / c)^2} - 1 \right) .$$

Eliminating $x(0)$ and squaring both sides gives eqn (3.25).

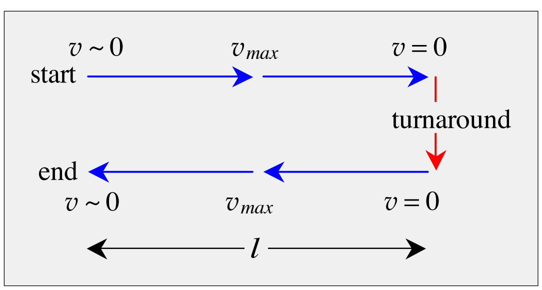
In summary, we have carried out a calculation whose result is that the squared interval $(x^1)^2 - (ct)^2$ is equal to $c^4(a^{1'})^{-2}$, for a constant value of proper acceleration $a^{1'}$. This is interesting but not surprising. Way back in the beginning, under the heading Galilean Relativity, we saw that in the Galileo-Newton theory the acceleration is the same in all inertial frames. Then later (Section 2.4) we found that the acceleration was the same everywhere on the hyperbola. Here we have arrived at essentially the same result. The proper acceleration experienced, say, by the woman in the pink car a few pages back is independent of the inertial frame from which she is observed. On the other hand, the acceleration $a' = du/dt = \gamma^{-3}a^{1'}$ depends on the inertial frame from which it is observed. As the world line approaches the light line, this acceleration approaches zero.

Example 3.4. Space Travel at Constant Acceleration



Let us return to the woman who was in the accelerating pink car. This time she is in a pink spaceship that has just taken off from earth and is sufficiently far from any gravitational field to justify ignoring it. Her spaceship is undergoing a constant acceleration of 10 m s^{-2}, so she weighs about the same in the spaceship as she would on earth. The magnitude of this acceleration remains constant. The acceleration can switch directions, but its magnitude must remain 10 m s^{-2}. After ten years of her proper time, τ, on the outward journey (v_{max} in the sketch below), the traveler tires of space travel. She reverses the direction of the acceleration, though its magnitude stays at 10 m s^{-2}.

Another ten years pass, and the spaceship velocity finally reaches zero. Here it turns around. After yet another ten years, the traveler is headed toward the earth with the same speed she had after the first ten years of the outgoing journey, v_{max}. She has now aged thirty years since the beginning of the trip. She reverses the direction of the acceleration to ensure a safe landing, which indeed occurs following the final ten-year slowdown period. The traveler aged a total of forty years during the full trip. How many years passed on earth during her trip? How far from earth did she venture at the furthest point, *i.e.*, the turnaround?



The best way to work through the numbers is by using the rapidity parameter ϕ, because all of its increments are additive: $e^{\phi_1 + \phi_2 + \phi_3 \cdots}$. The traveler is in her rest frame, except for infinitesimal increments, throughout her journey. She experiences acceleration of 10 m s^{-2} relative to a momentarily co-moving inertial frame. At any given moment, an infinitesimal proper time increment $d\tau$ is accompanied by a change of the rapidity from zero to $d\phi$. She then switches to another momentarily co-moving inertial frame and repeats the exercise. Someone who is recording the whole process from earth adds $d\phi$ to the ϕ that accumulated throughout the journey. From the relation $\tanh\phi = \beta$ (Section 2.4), we see that, as far as the traveler is concerned, each infinitesimal increment $d\tau$ yields $d\phi = d\beta = a^{1'} d\tau / c$, which integrates trivially:

$$\phi = \frac{a^{1'}\tau}{c} . \tag{1}$$

Thus, the rapidity ϕ is proportional to the traveler's proper time τ. Let us now express the distance traveled relative to an observer on earth. Use $\tanh\phi = \beta = dx / dct$ and $dt = \gamma d\tau = \cosh\phi \, d\tau$ to write

$$dx = c \tanh\phi \, dt \tag{2}$$

$$= c \tanh\phi \cosh\phi \, d\tau \tag{3}$$

$$= c \sinh\phi \, d\tau \tag{4}$$

$$= c \sinh\frac{a^{1'}\tau}{c} \, d\tau . \tag{5}$$

This integrates without ado, yielding

$$x - x(0) = c \int d\tau \, \sinh\frac{a^{1'}\tau}{c} \tag{6}$$

$$= \frac{c^2}{a^{1'}}\left(\cosh\frac{a^{1'}\tau}{c} - 1 \right) . \tag{7}$$

From an earlier figure we see that $x(0) = c^2 / a^{1'}$. Thus, we have a compact expression:

$$x(v_{max}) = \frac{c^2}{a^{1'}} \cosh\frac{a^{1'}\tau}{c} . \tag{8}$$

Now for numbers: The distance of furthest excursion is twice the value of $x(v_{max})$. The argument $a^{1'}\tau / c$ is equal to 10.5 for $\tau = 10$ years, and $c^2 / a^{1'}$ is equal to 9×10^{15}. Putting these numbers into eqn (8) and using 9.46×10^{15} meters in one light year, we find that the traveler reached an enormous distance at the turnaround: ~ 35,000 light years from earth.

Exercise: How many years passed on earth during the traveler's flight? Why is this amusing scenario not feasible?

3.7. Relativistic Kinetic Energy

In non-relativistic classical mechanics, the change of a particle's kinetic energy: $T = mv^2/2$, is equal to the integral of the three-force that the particle experiences projected onto the direction of its motion. What is referred to as relativistic kinetic energy is not $mv^2/2$, as in non-relativistic mechanics. Nonetheless, it depends on the particle's momentum, so it is still referred to as kinetic energy.

To obtain the relativistic version, we start with $F^\nu U_\nu = 0$. The time part is $F^0 U_0 = (dP^0/d\tau)\gamma c$, while the space part is $(d\vec{P}/d\tau)\gamma\vec{v}$. Writing this out yields

$$F^\nu U_\nu = 0 = \gamma\frac{dP^0}{d\tau}c - \gamma\frac{d\vec{P}}{d\tau}\cdot\vec{v}\,. \tag{i}$$

Using $P^0 = E/c$ and $\vec{P} = \gamma m\vec{v}$, and assuming that the particle starts from rest, integration becomes

$$\int_{E(v=0)}^{E} dE = \int_0^v d(\gamma m\vec{v})\cdot\vec{v}\,. \tag{ii}$$

Integrating by parts in the direction of motion incurs no loss of generality and yields for $T = E - E(v=0)$:

$$T = \gamma mv^2 - m\int_0^v \frac{vdv}{\sqrt{1-v^2/c^2}} = \gamma mv^2 + mc^2\sqrt{1-v^2/c^2}\,\Big|_0^v$$

$$= \gamma mv^2 + mc^2\sqrt{1-v^2/c^2} - mc^2\,. \tag{iii}$$

A little algebra (*i.e.*, $\gamma m(v^2 + c^2/\gamma^2) = \gamma mc^2(\cancel{v^2/}c^2 + 1 - \cancel{v^2/}c^2) = \gamma mc^2$) yields

$$T = \gamma mc^2 - mc^2\,. \tag{iv}$$

This identifies mc^2 as rest energy. Adding it to T yields the total energy

$$\boxed{E = \gamma mc^2\,.} \tag{v}$$

The contracted four-momentum: $P^\nu P_\nu = (mc)^2$ provides an alternate form

$$\boxed{E^2 = p^2c^2 + m^2c^4\,.} \tag{vi}$$

where p is the space part of the relativistic momentum. Note that when v is small eqn (iv) becomes

$$T + mc^2 \approx mc^2 + \tfrac{1}{2}mv^2\,. \tag{vii}$$

Again we see that mc^2 is the mass energy.

3.8. Summary

To see how the non-relativistic limit emerges from $E = \gamma mc^2$, expand the square root in the denominator of γ, then expand $(1-x)^{-1}$, and retain the leading terms:

$$\gamma mc^2 = \frac{mc^2}{\sqrt{1-v^2/c^2}} \tag{3.25}$$

$$= \frac{mc^2}{1-\frac{1}{2}v^2/c^2-\frac{1}{8}(v^2/c^2)^2\ldots} \tag{3.26}$$

$$= mc^2\left(1+\tfrac{1}{2}v^2/c^2+\tfrac{1}{8}(v^2/c^2)^2+\left(\tfrac{1}{2}v^2/c^2\right)^2\ldots\right) \tag{3.27}$$

$$= mc^2+\tfrac{1}{2}mv^2+\tfrac{3}{8}mv^2\frac{v^2}{c^2}\ \ldots \tag{3.28}$$

Alternatively, eqn (vi) on the previous page gives

$$E = mc^2\sqrt{1+(p/mc)^2} = mc^2\left(1+\tfrac{1}{2}\frac{p^2}{2m}-\tfrac{1}{8}\frac{p^4}{(mc)^4}\ldots\right) \tag{3.29}$$

Substitution of $p=\gamma mv$ recovers eqn (3.28) following some algebra.

The photon is now revealed through an intuitive (non-rigorous) approach. Specifically, the mass is made to vanish. Think of this as sneaking up on the result via the small mass of an electron neutrino (whose mc^2 value is ~ 1 eV) and then shrinking the mass all the way to zero. This is not quite right because a neutrino is a spin one-half fermion, but we shall do it anyway.

The four-momentum scalar product is

$$P^\nu P_\nu = (mc)^2 = (E/c)^2 - p^2, \tag{3.17}$$

where p^2 is the square of the space part of the four-momentum. As $m \to 0$, $P^\nu P_\nu = (mc)^2$ vanishes, leaving $p = E/c$ (positive square roots). The energy of the photon is $E = h\nu$, and consequently the photon momentum is $p = h\nu/c = h/\lambda$. This line of reasoning encouraged Louis de Broglie to surmise that the same relation holds for massive particles.

The results are summarized below.

$P^0 = \gamma mc = E/c$	$P^i = \gamma m\left(dx^i/dt\right)\ \left(i=1-3\right)$
$E = \gamma mc^2$	$E^2 = m^2c^4 + p^2c^2$

4. Four-Vectors

It is time to systematize notation. Until now, spacetime symbols and definitions have been introduced piecemeal and *ad hoc*: common usage (ct, x, y, z); subscripted (ct, x_1, x_2, x_3); and so on. This is inefficient, leaves loose ends, and it is hard to keep everything straight. The theory of special relativity is, among other things, an exercise in bookkeeping and geometry, and exploiting its geometric properties not only tidies nomenclature but also reveals the underlying physics in a way that is transparent and insightful. If you advance to general relativity (gravitation) or any other area that deals with curved space, you will find the mathematics introduced here helpful.

To begin, the variables ct, x, y, and z are replaced with their superscripted counterparts: x^0, x^1, x^2, and x^3, respectively. These are referred to as the contravariant components. In this section and the one that follows we will see why the term contravariant is used through an examination of the transformation properties of these components and the basis vectors that accompany them, as well as other geometric properties. I will stick to boosts.

Together with their basis vectors, the components x^0, x^1, x^2, and x^3 constitute a four-vector that differs qualitatively from the three-vectors of the 3D Euclidean space with which we have day-to-day familiarity. Likewise, relativistic momentum is a four-vector, and there are a number of other important four-vectors. For example, we frequently encounter the electromagnetic one whose contravariant components are: $A^\mu = (\phi, \vec{A}) = (A^0, A^1, A^2, A^3)$, where ϕ is the scalar electric potential and $\vec{A}$ is the magnetic vector potential. What distinguishes a mathematical object that is referred to as a four-vector is that it is form invariant under Lorentz transformation, as will be demonstrated. It is not at all the same as a vector in a 4D Euclidean space. The "variant" in covariant has obvious meaning. The "co" part stands for correlated, which makes sense insofar as how space and time change identity synchronously under boost transformation.

The Lorentz transformation that carries a system from unprimed to primed coordinates can be expressed as

$$x^{\mu'} = \Lambda^{\mu'}{}_{\nu}\, x^{\nu}\,, \tag{4.1}$$

which we have seen many times before in 2D form, for example,

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \underbrace{\begin{pmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{pmatrix}}_{\Lambda^{\mu'}{}_{\nu}} \begin{pmatrix} ct \\ x \end{pmatrix}.$$

As mentioned earlier, the widely accepted convention in place here is that Greek subscripts and superscripts are used with four-vectors, whereas Latin subscripts and superscripts are used with three-vectors. When you see an expression like x^ν, the free index ν indicates a four-vector. Likewise, the free index i in x^i indicates a three-vector.

The Lorentz transformation $\Lambda^{\mu'}{}_{\nu}$ can be assigned a 4×4 matrix representation. This is a matter of convenience. The needed manipulations can be carried out using nothing more than superscripted and subscripted components, as indicated in eqn (4.1),and sometimes it is not even possible to use matrices. For example, when dealing with tensors of rank greater than two, there are no matrix representations, so the math must be carried out using superscripts and subscripts. We will avoid such tensors, but it is something to keep in mind. However, matrices are relatively visual, and it is likely that you have familiarity with them, so they will be enlisted frequently.

To be careful with the matrix notation, the row index is placed to the left of the column index. How this works is seen in eqn (4.1), where the expression $\Lambda^{\mu'}{}_{\nu}\, x^{\nu}$ denotes the usual matrix multiplication, namely, the square matrix $\Lambda^{\mu'}{}_{\nu}$ multiplying from the left the column vector x^{ν}. The reason to be careful with the positioning of superscripts and subscripts is that when dealing with matrices and tensors it might not always be clear which index belongs with rows and which index belongs with columns. In the literature you will often see the superscript and subscript aligned vertically ($\Lambda^{\mu'}_{\nu}$). This is fine as long as no ambiguity arises, but it would not be a good idea here. As written, $\Lambda^{\mu'}{}_{\nu}$ is a single matrix element. However, common usage is that one often refers to a symbol like this as representing the full transformation.

Throughout the present and all following sections the Einstein summation convention is used. When the same index appears twice in an expression (once up, once down), it is to be summed over. For example, $x^{\mu}x_{\mu} = x^0x_0 + x^1x_1 + x^2x_2 + x^3x_3$. Finally, note that primes are placed on the numbers that label the basis for the new coordinate system, rather than on the components themselves. Lorentz transformation, in general, mixes components. For example, the new time coordinate contains the old time coordinate and the old space coordinate. It can be said that the basis is rotated, though a boost is obviously a special kind of rotation.[19]

The four-vector nature of the terms $x^{\mu'}$ and x^{ν} that appear in eqn (4.1) implies a coordinate origin (in 4D Minkowski space) that is analogous to the apex of the 3D light cone shown earlier. The vector extends from the origin to a spacetime point, with the contravariant components being the respective distances from the origin along the spacetime axes. To generalize this, a displacement four-vector $\Delta\vec{s}$ is defined in terms of the coordinate differences between two spacetime points separated by a straight line in the timelike region. In other words, the $x^{\mu'}$ and x^{ν} in eqn (4.1) are replaced with $\Delta x^{\mu'}$ and Δx^{ν}, respectively, yielding

$$\Delta x^{\mu'} = \Lambda^{\mu'}{}_{\nu}\, \Delta x^{\nu} . \tag{4.2}$$

The displacement four-vector $\Delta\vec{s} = \Delta x^{\mu}\, \vec{e}_{\mu}$ is given by

[19] The 2D transformation introduced in Section 1 mixes space and time. For the simple example of a boost in the x^1 direction, the boost changes the basis such that the new basis vectors contain the old ones: $\vec{e}_{1'}$ contains both $\vec{e}_1$ and $\vec{e}_0$.

$$\begin{pmatrix} \vec{e}_0 & \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{pmatrix} \begin{pmatrix} \Delta x^0 \\ \Delta x^1 \\ \Delta x^2 \\ \Delta x^3 \end{pmatrix} = \Delta x^0\,\vec{e}_0 + \Delta x^1\,\vec{e}_1 + \Delta x^2\,\vec{e}_2 + \Delta x^3\,\vec{e}_3 \tag{4.3}$$

The four-vector $\Delta\vec{s}$ is, of course, unaffected by a change of basis: $\Delta\vec{s} = \Delta x^{\mu}\,\vec{e}_{\mu} = \Delta x^{\mu'}\,\vec{e}_{\mu'}$. Though all of this appears harmless enough, a subtlety with the basis vectors soon will be revealed. For example, with $\Delta\vec{s}$ expressed as $\Delta x^{\mu}\,\vec{e}_{\mu}$, in order that $(\Delta\vec{s})^2$ returns the squared interval, the basis vectors must obey the relations: $\vec{e}_0\cdot\vec{e}_0 = 1$, whereas $\vec{e}_1\cdot\vec{e}_1 = \vec{e}_2\cdot\vec{e}_2 = \vec{e}_3\cdot\vec{e}_3 = -1$, and $\vec{e}_{\mu}\cdot\vec{e}_{\nu} = 0$ for $\mu \neq \nu$.

Equation (4.2) indicates how four-vector components behave under Lorentz transformation. They are said to be Lorentz covariant (form invariant): a four-vector remains a four-vector. The general expression for the matrix corresponding to $\Lambda^{\mu'}{}_{\nu}$ is

$$\Lambda^{\mu'}{}_{\nu} = \begin{pmatrix} \Lambda^{0'}{}_{0} & \Lambda^{0'}{}_{1} & \Lambda^{0'}{}_{2} & \Lambda^{0'}{}_{3} \\ \Lambda^{1'}{}_{0} & \Lambda^{1'}{}_{1} & \Lambda^{1'}{}_{2} & \Lambda^{1'}{}_{3} \\ \Lambda^{2'}{}_{0} & \Lambda^{2'}{}_{1} & \Lambda^{2'}{}_{2} & \Lambda^{2'}{}_{3} \\ \Lambda^{3'}{}_{0} & \Lambda^{3'}{}_{1} & \Lambda^{3'}{}_{2} & \Lambda^{3'}{}_{3} \end{pmatrix}. \tag{4.4}$$

The geometrical arrangement that was introduced in Section 1 [see Fig. 1.2 and eqns (1.2) and (1.3)] is as simple as it gets. In that case, only four of the 16 matrix elements indicated in eqn (4.4) have values other than 0 or 1, specifically, $\Lambda^{0'}{}_{0} = \Lambda^{1'}{}_{1} = \gamma$, and $\Lambda^{0'}{}_{1} = \Lambda^{1'}{}_{0} = -\beta\gamma$. The two other diagonal elements are each equal to 1 and the ten other off-diagonal elements are each equal to 0. It was also pointed out that the inverse transformation, in which an observer in the primed system makes measurements on the moving unprimed system, is obtained by exchanging the primed and unprimed labels and the direction of the velocity ($\beta \to -\beta$). In other words, the general expression for the matrix that describes this complementary transformation is

$$\Lambda^{\mu}{}_{\nu'} = \begin{pmatrix} \Lambda^{0}{}_{0'} & \Lambda^{0}{}_{1'} & \Lambda^{0}{}_{2'} & \Lambda^{0}{}_{3'} \\ \Lambda^{1}{}_{0'} & \Lambda^{1}{}_{1'} & \Lambda^{1}{}_{2'} & \Lambda^{1}{}_{3'} \\ \Lambda^{2}{}_{0'} & \Lambda^{2}{}_{1'} & \Lambda^{2}{}_{2'} & \Lambda^{2}{}_{3'} \\ \Lambda^{3}{}_{0'} & \Lambda^{3}{}_{1'} & \Lambda^{3}{}_{2'} & \Lambda^{3}{}_{3'} \end{pmatrix}. \tag{4.5}$$

The four non-trivial matrix elements are: $\Lambda^{0}{}_{0'} = \Lambda^{1}{}_{1'} = \gamma$, and $\Lambda^{0}{}_{1'} = \Lambda^{1}{}_{0'} = \beta\gamma$. Equations (4.4) and (4.5) become messy when explicit expressions for the general forms of the matrix elements are used, including 3D spatial rotation, as described in Section 6. However, it is rarely the case that the general expressions are used. It comes as no surprise that the product of the matrices in eqns (4.4) and (4.5) is the unit matrix, as going from an unprimed frame to a primed frame and then back to the unprimed frame is the equivalent of nothing happening, and likewise for rotations in opposite senses. The math is verified with a line or two of algebra for the case of only four non-trivial matrix elements. To verify it mathematically for the general case, start with eqn (4.2):

$$\Delta x^{\mu'} = \Lambda^{\mu'}{}_{\nu}\, \Delta x^{\nu} \,, \tag{4.2}$$

and use the fact that $\Delta x^{\nu} = \Lambda^{\nu}{}_{\varepsilon'}\, \Delta x^{\varepsilon'}$ to write

$$\Delta x^{\mu'} = \Lambda^{\mu'}{}_{\nu}\, \Lambda^{\nu}{}_{\varepsilon'}\, \Delta x^{\varepsilon'} \,. \tag{4.6}$$

This is only satisfied if $\mu' = \varepsilon'$. Thus, we have established the useful relationship

$$\Lambda^{\mu'}{}_{\nu}\, \Lambda^{\nu}{}_{\varepsilon'} = \delta^{\mu'}_{\varepsilon'} \,, \tag{4.7}$$

where $\delta^{\mu'}_{\varepsilon'}$ is the Kronecker delta.

Let us now examine the transformation properties of the contravariant components and their corresponding covariant basis vectors. The contravariant components are what we have come to accept as the standard components of a vector, albeit extended to 4D spacetime. The vector $\Delta\vec{s} = \Delta x^{\mu}\vec{e}_{\mu}$ is the same in any reference frame, so $\Delta x^{\mu}\vec{e}_{\mu}$ can also be written $\Delta x^{\mu'}\vec{e}_{\mu'}$. In addition, we know that the components transform according to eqn (4.2), *i.e.*, $\Delta x^{\mu'} = \Lambda^{\mu'}{}_{\nu}\, \Delta x^{\nu}$. But how do the basis vectors transform? To answer this, use the above facts to write

$$\Delta\vec{s} = \Delta x^{\mu'}\vec{e}_{\mu'} \tag{4.8}$$

$$= \Lambda^{\mu'}{}_{\nu}\, \Delta x^{\nu}\, \vec{e}_{\mu'} \tag{4.9}$$

$$= \Delta x^{\nu} \Lambda^{\mu'}{}_{\nu}\, \vec{e}_{\mu'} \,. \tag{4.10}$$

Because $\Delta\vec{s}$ is the same in all frames, this must be equivalent to

$$\Delta\vec{s} = \Delta x^{\nu}\, \vec{e}_{\nu} \,. \tag{4.11}$$

In comparing eqns (4.10) and (4.11), we see that

$$\vec{e}_{\nu} = \Lambda^{\mu'}{}_{\nu}\, \vec{e}_{\mu'} \,. \tag{4.12}$$

Now multiply both sides of eqn (4.12) by $\Lambda^{\nu}{}_{\varepsilon'}$ and use eqn (4.7):[20]

[20] Notice that Einstein summation does not respect the order in which the contravariant and covariant indices appear, whereas matrix multiplication is non-commutative.

$$\Lambda^{\nu}{}_{\varepsilon'}\,\vec{e}_{\nu} = \Lambda^{\nu}{}_{\varepsilon'}\,\Lambda^{\mu'}{}_{\nu}\,\vec{e}_{\mu'} \tag{4.13}$$

$$= \delta^{\mu'}_{\varepsilon'}\,\vec{e}_{\mu'}\,. \tag{4.14}$$

Thus,

$$\vec{e}_{\mu'} = \Lambda^{\nu}{}_{\mu'}\,\vec{e}_{\nu}\,. \tag{4.15}$$

From the above manipulations, we see that the contravariant components transform according to $\Delta x^{\mu'} = \Lambda^{\mu'}{}_{\nu}\Delta x^{\nu}$, whereas the corresponding covariant basis vectors transform according to $\vec{e}_{\mu'} = \Lambda^{\nu}{}_{\mu'}\,\vec{e}_{\nu}$.

$$\boxed{\begin{gathered}\Delta x^{\mu'} = \Lambda^{\mu'}{}_{\nu}\Delta x^{\nu}\\ \vec{e}_{\mu'} = \Lambda^{\nu}{}_{\mu'}\,\vec{e}_{\nu}\end{gathered}}$$

The components and basis vectors transform oppositely, in the sense that the transformations $\Lambda^{\mu'}{}_{\nu}$ and $\Lambda^{\nu}{}_{\mu'}$ are the inverses of one another, *i.e.*, $\Lambda^{\mu'}{}_{\nu}\,\Lambda^{\nu}{}_{\varepsilon'} = \delta^{\mu'}_{\varepsilon'}$. Some people say that this explains why the term contravariant is used. The transformation properties of the contravariant components ensure that $\Delta\vec{s}$ is the same in all reference frames.

Contravariant and Covariant Components and Basis Vectors

We now turn to the geometric interpretation that goes hand in hand with the definitions and algebraic manipulations of contravariant and covariant quantities presented in the previous section. We have seen that relativity deals with non-orthogonal reference frames, as manifest for example in the Minkowski space diagrams that relate unprimed and primed reference frames. In addition, an interesting metric has emerged that returns the squared interval, ds^2, and like quantities for other four-vectors. Such mathematical tools are essential if we are to deal efficiently with transformations between non-orthogonal reference frames, as well as compute four-vector scalar products, four divergences, and analogous operations on tensors of higher rank. The goal of the present section is to provide further insight into how this machinery works.

Several mathematical relations of purely geometric nature are derived here. They are then used with Lorentz transformation, and we will see that things fall neatly into place. Later, we will encounter a four-vector synchrony between special relativity and electrodynamics that is, as the cliché goes, "too good to be true." Likewise, we will see that the gauge principle that underlies quantum electrodynamics (QED) links quantum mechanics and electrodynamics in a way that again is too good to be true. It also links the weak and strong nuclear forces to quantum mechanics (the standard model of physics), but such is beyond the scope of this course.

Make no mistake about it: The algebra and strategies for dealing with contravariant and covariant quantities discussed in this and the previous section will prove indispensable in later chapters.

To begin, consider the 2D systems shown in Fig. 4.1: Entry (a) is the standard Cartesian frame. (b) In this skewed-axis frame, axis-parallel projections are used to obtain a^1 and a^2, which are referred to as the contravariant components. They are *defined* as axis-parallel projections. Their covariant basis vectors are labeled $\hat{g}_1$ and $\hat{g}_2$ to distinguish them from the basis vectors $\hat{e}_1$ and $\hat{e}_2$ ($\hat{e}_x$ and $\hat{e}_y$) of the Cartesian frame.

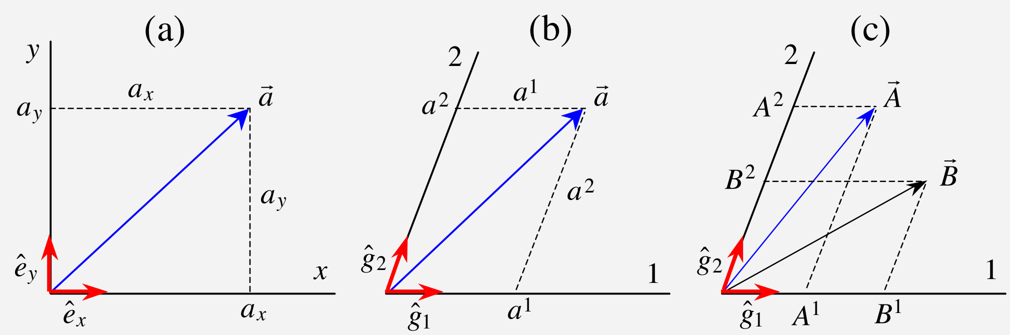


Figure 4.1. (a) Cartesian frame. (b) Skewed-axes frame showing axis-parallel projections. Superscripts and subscripts denote contravariant and covariant quantities, respectively. (c) Two vectors in the skewed-axes frame: Their inner product can be expressed in terms of contravariant components and covariant basis vectors [*vide infra*, eqns (2) and (3)].

Next, consider the scalar product of two vectors $\vec{A}$ and $\vec{B}$ in a 2D space. In a frame with orthogonal basis vectors, this is written: $\vec{A}\cdot\vec{B} = A^i B^j \hat{e}_i \cdot \hat{e}_j$, with the basis vectors obeying $\hat{e}_i \cdot \hat{e}_j = \delta_{ij}$. If the axes are not orthogonal but skewed, as in Fig. 1(c), the same dot product operation yields

$$\left(A^1\hat{g}_1 + A^2\hat{g}_2\right)\cdot\left(B^1\hat{g}_1 + B^2\hat{g}_2\right) = A^1B^1\hat{g}_1\cdot\hat{g}_1 + A^2B^2\hat{g}_2\cdot\hat{g}_2 + A^1B^2\hat{g}_1\cdot\hat{g}_2 + A^2B^1\hat{g}_2\cdot\hat{g}_1 \tag{1}$$

If we define the $\hat{g}_i$ to be unit vectors, this becomes

$$\vec{A}\cdot\vec{B} = A^1\underbrace{\left(B^1 + B^2\hat{g}_1\cdot\hat{g}_2\right)}_{B_1} + A^2\underbrace{\left(B^2 + B^1\hat{g}_2\cdot\hat{g}_1\right)}_{B_2} \tag{2}$$

$$= A^1B_1 + A^2B_2\,. \tag{3}$$

The B_i in the above expression are referred to as the covariant components.

The geometric picture associated with the above equations is shown in Fig. 2. Let us focus on the B_1 component, as the B_2 component follows for like reason. In terms of contravariant components, it is written: $B_1 = B^1 + B^2 \hat{g}_1 \cdot \hat{g}_2 = B^1 + B^2 \cos\theta$. You can see from Fig. 2 and the above equation that the length of B_1 is given by the perpendicular projection of the vector $\vec{B}$ onto the 1-axis. Be careful, however. Figure 2 gives the lengths B_1 and B_2, but we do not yet have the associated basis vectors $\hat{g}^1$ and $\hat{g}^2$, and they certainly differ from $\hat{g}_1$ and $\hat{g}_2$. In other words, the quantities B_1 and B_2 are *not* aligned with the basis vectors $\hat{g}_1$ and $\hat{g}_2$, despite the fact that they are projections onto the 1 and 2 axes, respectively. This is subtle. Rather, they are aligned with $\hat{g}^1$ and $\hat{g}^2$, which are indicated in Fig. 3.

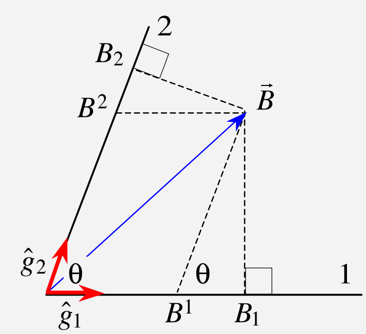


Figure 2. Covariant components are perpendicular projections onto the 1 and 2 axes.

In general, the covariant components can be obtained from the contravariant components through a straightforward operation. In the present example, this operation can be expressed in its most general form as

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} \hat{g}_1 \cdot \hat{g}_1 & \hat{g}_1 \cdot \hat{g}_2 \\ \hat{g}_2 \cdot \hat{g}_1 & \hat{g}_2 \cdot \hat{g}_2 \end{pmatrix} \begin{pmatrix} B^1 \\ B^2 \end{pmatrix}. \tag{4}$$

It is understood that higher dimensional spaces follow suit. The 2×2 matrix in eqn (4), or its equivalent, is referred to as the metric. The metric need not be expressed as a matrix, but I thought this would be familiar to most of you. The metric for the orthogonal axis system in Fig. 1(a) is diag(1,1), so we need not pay attention to it, but for skewed axes it is important. To learn more about the covariant vector components, and particularly their contravariant basis vectors, we write $\vec{A} \cdot \vec{B}$ as

$$\left(A^1 \hat{g}_1 + A^2 \hat{g}_2\right) \cdot \left(B_1 \hat{g}^1 + B_2 \hat{g}^2\right) = A^1 B_1 \hat{g}_1 \cdot \hat{g}^1 + A^2 B_2 \hat{g}_2 \cdot \hat{g}^2 + A^1 B_2 \hat{g}_1 \cdot \hat{g}^2 + A^2 B_1 \hat{g}_2 \cdot \hat{g}^1 \tag{5}$$

Referring to eqn (3), we see that the right hand side of eqn (5) must be equal to $A^i B_i$. This requires that

$$\hat{g}_i \cdot \hat{g}^j = \delta_i{}^j . \tag{6}$$

We now have the desired geometric interpretation, which is indicated in Fig. 3. The covariant unit basis vectors $\hat{g}_1$ and $\hat{g}_2$ are orthogonal, respectively, to the (blue) contravariant basis vectors $\hat{g}^2$ and $\hat{g}^1$. The lengths of the contravariant basis vectors are established because their projections onto the covariant basis vectors, by definition, have unit length.

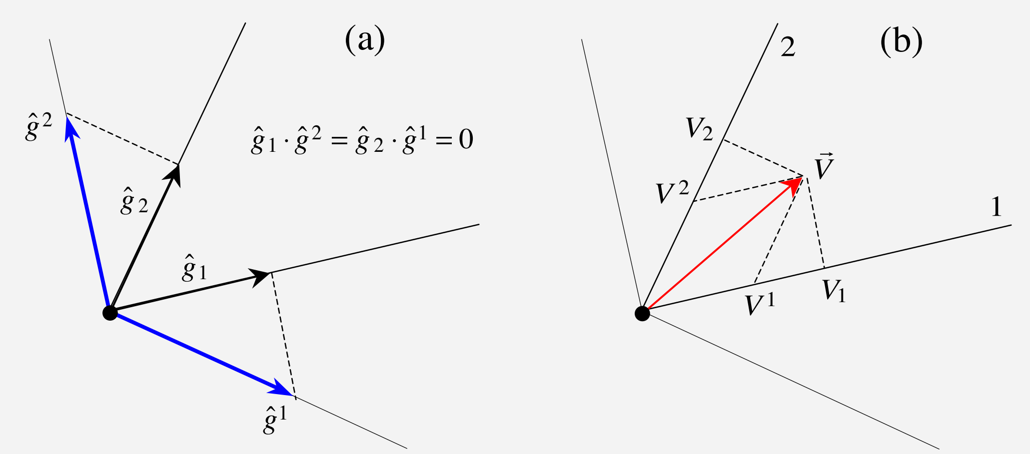


Figure 3. (a) The basis vectors obey the relation: $\hat{g}_i \cdot \hat{g}^j = \delta_i{}^j$, with the magnitudes of the covariant basis vectors $\hat{g}_i$ defined as unity. (b) The contravariant and covariant components are parallel and perpendicular projections, respectively.

The vector components also have been obtained earlier. The parallel-axis projections give the contravariant components V^1 and V^2, whereas the covariant components V_1 and V_2 are the perpendicular projections onto the axes. As mentioned earlier, the covariant components are taken from the axes of the covariant basis, rather than the axes of the contravariant basis. This can be confusing, but if you go through it a number of times it will sink in.

Let us now advance to special relativity and the Lorentz transformation. This time we shall go from unprimed Cartesian axes to primed skewed axes. Let us also pretend that we do not already know the answers. The first odd thing we notice is that the unit basis vectors in the unprimed system appear to get bigger when they are transformed into the primed system, as seen for example with the transformation:

$$\begin{pmatrix} \hat{g}_{0'} \\ \hat{g}_{1'} \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} \hat{g}_0 \\ \hat{g}_1 \end{pmatrix}. \tag{7}$$

When $\hat{g}_{0'} \cdot \hat{g}_{0'}$ is computed using $\hat{g}_{0'} = \gamma \hat{g}_0 - \beta\gamma \hat{g}_1$, we get

$$\hat{g}_{0'} \cdot \hat{g}_{0'} = (\gamma \hat{g}_0 - \beta\gamma \hat{g}_1) \cdot (\gamma \hat{g}_0 - \beta\gamma \hat{g}_1) \tag{8}$$

$$= \gamma^2 \hat{g}_0 \cdot \hat{g}_0 + \gamma^2 \beta^2 \hat{g}_1 \cdot \hat{g}_1 \tag{9}$$

$$= \gamma^2 (1 + \beta^2). \tag{10}$$

The value obtained, $\gamma^2(1+\beta^2)$, clearly exceeds unity; and likewise for $\hat{g}_{1'}$. The same occurs with other vector components such as ct and x.

The problem lies with the choice of metric. Computing $\hat{g}_{0'} \cdot \hat{g}_{0'}$ this way implies a metric of the form given by eqn (4), with standard (positive real) basis vectors. This works in many physical situations, but not here. In order to recover the squared interval of special relativity, it is necessary to use the Minkowski metric, which has the interesting property that the basis vectors for the spatial components are imaginary. Recall that we have opted for the diag(+ – – –) option rather than diag(– + + +). Earlier it was pointed out that Poincaré's original suggestion of using *ict* for the time coordinate has been replaced by transferring the imaginary unit to the basis vectors. This is what we are doing here.

All of this business about an imaginary coordinate, and transferring the *i* in *ict* to the basis vectors, sounds more sophisticated than it is. Poincaré just invoked *ict* to get the job done, not through some deep physical reasoning. The physics of classical special relativity requires a minus sign in the squared interval, and *ict* works. Thus, transferring the *i* to the basis vectors is exchanging one *ad hoc* maneuver for another. In fact, no complex quantities should appear in the outcome of any classical physics calculation. The bottom line is that you need only define $\hat{g}_i \cdot \hat{g}_j = -\delta_i{}^j$. The physics requires this, and that is all the justification that is needed.

The matrix representation of the diag(+ – – –) Minkowski metric, $\eta_{\mu\nu}$, is

$$\eta_{\mu\nu} = \begin{pmatrix} \hat{g}_0 \cdot \hat{g}_0 & 0 & 0 & 0 \\ 0 & \hat{g}_1 \cdot \hat{g}_1 & 0 & 0 \\ 0 & 0 & \hat{g}_2 \cdot \hat{g}_2 & 0 \\ 0 & 0 & 0 & \hat{g}_3 \cdot \hat{g}_3 \end{pmatrix} \tag{10}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{11}$$

With the Minkowski metric everything falls into place. If we repeat the calculation of $\hat{g}_{0'} \cdot \hat{g}_{0'}$, but this time with the Minkowski metric, we get

$$[\gamma, -\beta\gamma]\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} \gamma \\ -\beta\gamma \end{bmatrix} = \gamma^2\left(1-\beta^2\right) = 1\,. \tag{12}$$

Likewise, if we compute $\hat{g}_{1'} \cdot \hat{g}_{1'}$ using $\hat{g}_{1'} = -\beta\gamma\hat{g}_0 + \gamma\hat{g}_1$ we find

$$[-\beta\gamma, \gamma]\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} -\beta\gamma \\ \gamma \end{bmatrix} = \gamma^2\left(\beta^2 - 1\right) = -1\,. \tag{13}$$

Other combinations do equally well. You should work though these. Let us now look further into the metric tensor.

5. Metric Tensor

The mathematics that underlies special relativity is geometric in nature. It deals with non-orthogonal reference frames and transformations between contravariant and covariant components and basis vectors. A central mathematical object is a metric tensor that defines the properties of the space, illustrates relationships, and facilitates their manipulations. The present section provides an overview. A good account is given in Kusse and Westwig [16] in the chapter: Tensors in Non-orthogonal Coordinate Systems. There is also an earlier chapter on tensors and other relevant material.

The contravariant components Δx^0 , Δx^1 , Δx^2 , and Δx^3 can be used to calculate the squared interval Δs^2 . However, the math is subtle. For example, if we form the vector: $\Delta x^0 \vec{e}_0 + \Delta x^1 \vec{e}_1 + \Delta x^2 \vec{e}_2 + \Delta x^3 \vec{e}_3$ and take the usual Euclidean dot product of this vector with itself, a quantity is obtained that is not Lorentz invariant: $(\Delta x^0)^2 + (\Delta x^1)^2 + (\Delta x^2)^2 + (\Delta x^3)^2$. Clearly, this quantity does not match the Lorentz invariant squared interval Δs^2 or its negative, $-\Delta s^2$. It was for this reason that Poincaré introduced *ict*. Squaring *ict* gives $-(ct)^2$, and a Lorentz invariant quantity is obtained.

The modern way is to transfer the imaginary unit i to the basis. In other words, introduce imaginary basis vectors according to either of two options: (1) $\vec{e}_0$, $i\vec{e}_1$, $i\vec{e}_2$, $i\vec{e}_3$, or, alternatively, (2) $i\vec{e}_0 , \vec{e}_1 , \vec{e}_2 , \vec{e}_3$. Each of these options works and is in common use. We will use the former, so instead of $\vec{e}_1 \cdot \vec{e}_1 = 1$, we have $i\vec{e}_1 \cdot i\vec{e}_1 = -1$, and likewise $i\vec{e}_2 \cdot i\vec{e}_2 = -1$, and $i\vec{e}_3 \cdot i\vec{e}_3 = -1$. As mentioned earlier, an alternate (better) way of stating this is simply to define $\vec{e}_0 \cdot \vec{e}_0 = 1$, $\vec{e}_1 \cdot \vec{e}_1 = \vec{e}_2 \cdot \vec{e}_2 = \vec{e}_3 \cdot \vec{e}_3 = -1$, and $\vec{e}_\mu \cdot \vec{e}_\nu = 0$ for $\mu \neq \nu$.

This yields the metric tensor on Minkowski space: $\eta_{\mu\nu} = \vec{e}_\mu \cdot \vec{e}_\nu$, with the Minkowski signature, denoted diag$(+---)$. Again, keep in mind that diag$(-+++)$ is also acceptable. This approach enables Δs^2 to be calculated in a way that is analogous to the standard dot product on Euclidean space. To see how this works, consider the matrix representation of the metric tensor $\eta_{\mu\nu}$:

$$\eta_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{5.1}$$

When this acts on the contravariant components arranged in a column vector, the needed sign is obtained:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \Delta x^0 \\ \Delta x^1 \\ \Delta x^2 \\ \Delta x^3 \end{pmatrix} = \begin{pmatrix} \Delta x^0 \\ -\Delta x^1 \\ -\Delta x^2 \\ -\Delta x^3 \end{pmatrix} \tag{5.2}$$

$$= \begin{pmatrix} \Delta x_0 \\ \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \end{pmatrix}. \tag{5.3}$$

In this case, the required metric tensor is $\eta_{\mu\nu}$. It has two lower indices, making it the covariant representation. When it acts on the contravariant components Δx^ν, their indices are lowered and the covariant components Δx_ν are obtained. Thus, we see that $\Delta x_0 = \Delta x^0$, $\Delta x_1 = -\Delta x^1$, $\Delta x_2 = -\Delta x^2$, $\Delta x_3 = -\Delta x^3$. The entries in the column vectors on the right hand sides of eqns (5.2) and (5.3) are the covariant components.

Acting on the covariant components with $\eta^{\mu\nu}$ returns the column vector to its original contravariant representation. Thus, $\eta^{\mu\nu}$, which raises covariant components to contravariant ones, has the same form as eqn (5.1). The matrix representations of $\eta_{\mu\nu}$ and $\eta^{\mu\nu}$ are the same, though it is understood that, respectively, they change contravariant to covariant and *vice versa*.

The scalar Δs^2 is obtained by multiplying the right side of eqn (5.2) from the left by the row matrix ($\Delta x^0\ \Delta x^1\ \Delta x^2\ \Delta x^3$), thus obtaining $(\Delta x^0)^2 - (\Delta x^1)^2 - (\Delta x^2)^2 - (\Delta x^3)^2$. There are a number of worthwhile texts that deal with such manipulations. If you are ambitious, the excellent book by Doran and Lasenby [21] offers the enlightened view afforded by geometric algebra. The box below lists a few manipulations that are germane to the above discussion.

A Few Manipulations Using $\eta_{\mu\nu}$ and $\eta^{\mu\nu}$

For the sake of brevity, nearly all of the Δ's are suppressed (taken as understood).

Lowering the indices of x^ν with the metric yields the covariant components:

$$\eta_{\mu\nu} x^\nu = x_\mu \Rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} = \begin{pmatrix} x^0 \\ -x^1 \\ -x^2 \\ -x^3 \end{pmatrix} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Thus, the scalar Δs^2 is given by $x^\mu x_\mu = x^\mu \eta_{\mu\nu} x^\nu = (x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2$. This is expressed in matrix form as:

$$\begin{pmatrix} x^0 & x^1 & x^2 & x^3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} = \underbrace{\begin{pmatrix} x^0 & x^1 & x^2 & x^3 \end{pmatrix} \begin{pmatrix} x^0 \\ -x^1 \\ -x^2 \\ -x^3 \end{pmatrix}}_{(x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2}$$

The same result is obtained by working from the other direction. Namely, use $x^\mu = \eta^{\mu\nu} x_\nu$ and form $\Delta s^2 = x_\mu x^\mu = x_\mu \eta^{\mu\nu} x_\nu = (x_0)^2 - (x_1)^2 - (x_2)^2 - (x_3)^2$. Note that this is equivalent to $(x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2$.

General relativity is based on the equivalence between gravity and the curvature of space brought about by mass. The metric in this case is not as simple as that of Minkowski space. Dynamics follow from operations on the metric tensor. For example, Christoffel symbols (connection coefficients) are derivatives of the metric tensor with respect to coordinates on the four-manifold. When a space is orthogonal, the metric has only diagonal matrix elements, and when it is possible to transform the coordinates to ones in which the elements of the metric are constants, the space is flat. Thus, the four-space of special relativity is flat. When the metric has elements that vary from one point to another and it is not possible to remove this variation by a transformation, the space is intrinsically curved, a simple example being the surface of a sphere.

The case of spherical coordinates serves to illustrate both curved and flat spaces. The directions $\hat{r}$, $\hat{\theta}$, and $\hat{\phi}$ are mutually orthogonal, and dr, $d\theta$, and $d\phi$ are the respective increments in these directions. These are not lengths, but increments in the variables r, θ, and ϕ. To obtain the squared differential length, use the metric tensor

$$g = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix}, \tag{5.4}$$

which yields

$$\begin{pmatrix} dr & d\theta & d\phi \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix} \begin{pmatrix} dr \\ d\theta \\ d\phi \end{pmatrix} = (dr)^2 + (rd\theta)^2 + (r\sin\theta\, d\phi)^2 . \tag{5.5}$$

The spherical coordinate system can be transformed to one of Cartesian coordinates, so the space is inherently flat despite the fact that the matrix elements are not constants. The surface of a sphere is a different matter. There is no way to transform coordinates on its surface to a flat representation.

6. Adding 3D Rotation

The Lorentz transformation is obtained here using general arguments and group theoretical considerations. The approach makes clear what is included versus what is not. So far, we have dealt almost exclusively with boosts because they mix space and time coordinates and, at least in classical physics, this is the most important relativistic effect. However, we will see that 3D spatial rotations are integral to special relativity. For example, if motion takes place in a direction that is inconvenient mathematically, the frame can be rotated such that a convenient alignment is achieved. After accounting for relativistic boost(s), the frame can then be rotated back, or elsewhere, if one wishes.

Parity and time reversal symmetries are also part of the full Lorentz transformation, but they are not part of the special relativity transformations under consideration here, because they are discrete, *i.e.*, they do not vary continuously on spacetime. Parity and time reversal are germane to the quantum case. Overall translation ($\Delta x^{\mu'} = \Lambda^{\mu'}{}_{\nu} \Delta x^{\nu} + a^{\mu'}$) is not included, but is an ingredient of the (more encompassing) Poincaré group, which is also referred to as the inhomogeneous Lorentz group.

In matrix form the invariance of Δs^2 is expressed as: $\Delta s^2 = \Delta x^{\mathrm{T}} \eta \, \Delta x = \Delta x'^{\mathrm{T}} \eta \, \Delta x'$, where T denotes transpose and η is the spacetime metric introduced in the last section It is understood that $\eta_{\mu\nu}$ is used in combination with contravariant components Δx^{ν}, while $\eta^{\mu\nu}$ is used in combination with covariant components Δx_{ν}. Because $\Delta x' = \Lambda \Delta x$, the expression for Δs^2 becomes: $\Delta s^2 = (\Delta x^{\mathrm{T}} \Lambda^{\mathrm{T}}) \eta \, (\Lambda \Delta x)$. Consequently, the Lorentz transformation Λ obeys the equation:[21]

$$\eta = \Lambda^{\mathrm{T}} \eta \, \Lambda \, . \tag{6.1}$$

Matrices that satisfy this equation are sought.[22]

Recall that the metric signature we have chosen places +1 at the (0, 0) position, –1 at the three other diagonal positions, and zeros at all of the off-diagonal positions. The transformations for the 3×3 matrix having –1 diagonal entries are referred to as orthogonal, as the determinant is (effectively!) equal to +1. Specifically, note than this 3×3 sub-matrix is equal to –1 times a unit matrix. The overall minus is unimportant and it is therefore ignored, for example, it does not arise when the other metric signature is chosen, diag(–+++).

This orthogonal transformation includes rotations and parity. However, parity must be excluded, as it is a discrete symmetry, and the Lorentz transformations of interest here must vary continuously about the identity. They have infinitesimal generators [8]. Likewise, time-reversal symmetry must be excluded. It is also a discrete symmetry: a par-

---

[21] In component form eqn (6.1) is: $\eta_{\varepsilon\sigma} = \Lambda^{\rho'}{}_{\varepsilon} \, \eta_{\rho'\mu'} \, \Lambda^{\mu'}{}_{\sigma} = \Lambda^{\rho'}{}_{\varepsilon} \Lambda^{\mu'}{}_{\sigma} \, \eta_{\rho'\mu'}$. Notice that order counts when multiplying matrices but not when contracting components.

[22] This defines the Poincaré group. We are interested in its Lorentz subgroup.

ity operation in Minkowski space, so to speak. The manner in which rotations enter the Lorentz transformation is obvious from 3D rotations. For example, rotation of the $\hat{e}_1 / \hat{e}_2$ axes by an angle θ about $\hat{e}_3$ is given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & \sin\theta & 0 \\ 0 & -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{6.2}$$

and likewise for rotations about the $\hat{e}_1$ and $\hat{e}_2$ axes. On the other hand, boosts involve time. For motion along $\hat{e}_1$, *i.e.*, which involves x^0 and x^1, we have

$$\begin{pmatrix} \cosh\phi & -\sinh\phi & 0 & 0 \\ -\sinh\phi & \cosh\phi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{6.3}$$

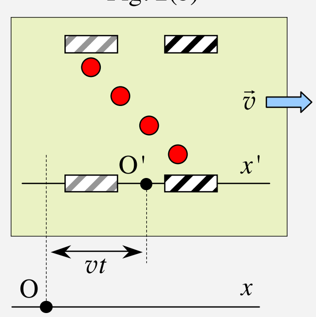
Recall that ϕ is specified in terms of β and γ. To see how this works, write the 2×2 part:

$$\begin{pmatrix} x^{0'} \\ x^{1'} \end{pmatrix} = \begin{pmatrix} \cosh\phi & -\sinh\phi \\ -\sinh\phi & \cosh\phi \end{pmatrix} \begin{pmatrix} x^0 \\ x^1 \end{pmatrix}. \tag{6.4}$$

The figure on the right shows that $x^{1'} = 0$ when $x^1 = vt$. Using this with eqn (6.4) yields: $ct \sinh\phi = vt \cosh\phi$, or $\tanh\phi = \beta$. Thus, $\cosh^2\phi = (1-\beta^2)^{-1} = \gamma^2$, recovering eqn (1.3):

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix}. \tag{1.3}$$

Fig. 2(b)



Bibliography and References

1. A. Einstein, Zur Elektrodynamik der bewegten Körper, Annalen der Physik **17**, 891-921 (1905). You can find English translations online.
2. H. Poincaré, La mesure du temps, Revue de métaphysique et de morale **6**, 1-7 (1898).
3. H. Poincaré, La Science et l'hypothèse, Edition Flammarion, Paris, 1902.
4. H. Poincaré, Sur la dynamique de l'électron, Comptes Rendus Acad. Sci. Paris **140**, 1504-1508 (1905). A follow-on paper is: H. Poincaré, Sur la dynamique de l'électron. Rendiconti del Circolo Matematico di Palermo **21**, 129-175, received 23 July, 1905, published January 1906.
5. E. F. Taylor and J. A. Wheeler, Spacetime Physics (W. H. Freeman and Company, San Francisco, 1992). This legendary text was first published in 1963. It contains an enormous number of great problems and examples. The math level is exceptionally modest. The authors stick to the basics, encouraging the reader to come to grips with the manifestations and consequences of relativity by developing intuition.
6. W. Rindler, Relativity: Special, General, and Cosmological, 2nd Edition (Oxford University Press, New York, 2006). An outstanding book that offers deep physical insight despite a few dated ideas and explanations. It is an excellent complement to Taylor and Wheeler. First read Taylor and Wheeler, and then read Rindler.
7. R. Ferraro, Introduction to Special and General Relativity (Springer, Inc., New York, 2007).
8. C. Schiller, Motion Mountain, CreateSpace Independent Publishing Platform; 25.6 edition (December 7, 2013).
9. N. M. J. Woodhouse, Special Relativity (Springer-Verlag, Undergraduate Mathematics Series, London, 2003). This book is aimed at undergraduates; accordingly, the math level is modest. Introductory material is especially well done.
10. D. Bohm, The Special Theory of Relativity (Routledge, New York, 1965). There is rather little math in this somewhat philosophical book.
11. A. P. French, Special Relativity (CRC Press, New York, 1968).
12. A. Harpaz, Relativity Theory (A. K. Peters, Wellesley, Massachusetts, 1993).
13. C. Misner, K. Thorne, and J. A. Wheeler, Gravitation (W. H. Freeman and Company, San Francisco, 1973). Brace yourself. This 1300 page tome is considered the "industry standard." Hopefully it will not take you as long to read and understand this book as it took the authors to write it (decades).
14. S. Carroll, Spacetime and Geometry: An Introduction to General Relativity (Addison Wesley, New York, 2004). This is a beautifully written text, though not for someone just starting out on the subject relativity.
15. J. Norton, University of Pittsburgh, Department of History and Philosophy of Science www.pitt.edu/~jdnorton/teaching/HPS_0410/chapters/Special_relativity_rel_sim

16. B. Kusse and E. Westwig, Mathematical Physics: Applied Mathematics for Scientists and Engineers (Wiley, New York, 2010). This is an excellent all-around reference. It explains nicely the geometry of covariant and contravariant components and basis vectors together with associated transformations.
17. G. B. Arfken and H. J. Weber, Mathematical Methods for Physicists, 5th Edition (Academic Press, New York, 2001). There are many editions of this book, but they hardly differ in content.
18. J. D. Jackson, Classical Electrodynamics, 3rd Edition (Wiley, New York, 1999).
19. E. M. Purcell, Electricity and Magnetism, 2nd Edition (McGraw-Hill, New York 1985).
20. J. D. Cresser, Lecture Notes on Special Relativity, Macquarie University Physics Department (2005).
21. C. Doran and A. Lasenby, Geometric Algebra for Physicists (Cambridge University Press, Cambridge, 2003).
22. D. Hestenes, New Foundations for Classical Mechanics, 2nd Edition (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999).
23. F. Schwabl, Advanced Quantum Mechanics, 4th Edition (Springer, Berlin, 2008).
24. J. A. Rhodes and M. D. Semon, Relativistic Velocity Space, Wigner Rotation, and Thomas Precession, Am. J. Phys. **72**, 943-960 (2004).
25. M. Schwartz, Principles of Electrodynamics (McGraw-Hill, New York, 1972).