CHAPTER 1

Time, Space, and Space-Time

I have seen many books which have objected to [Euclid's fifth postulate], among the earlier ones Heron and Autocus (Autolycus),and the later ones Al-Khazen, Al-Sheni, Al-Neyrizi, etc. None has given a proof. Then I have seen the book of Ibn Haytham, God bless his soul, called the solution of doubt in Chapter One. This postulate among other things was accepted without proof. There are many other things which are foreign to this field, such as:

If a straight line segment moves so that it remains perpendicular to a given line, and one end of it remains on the given line, then the other end of it draws a parallel.

There are many things wrong here. How could a proof be based on this idea? How could geometry and motion be connected?

Omar Khayyám (1048–1131). See [**1**], p. 277.

1. Simultaneity and Sequentiality

Omar Khayyám's criticism of the principle behind an attempted proof of the parallel postulate by ibn al-Haytham has some features in common with a basic principle of the special theory of relativity. In the attempted proof, one line segment is pictured as sliding along another one so as to be always perpendicular to it. Omar Khayyám merely asked how a line segment[1] could move. He was willing to grant that a point could move, and that all the points in a collection could move individually. But, he thought, if they move, they are no longer the same points they were before. How can we be sure that the relations among them will be preserved in such a way that they continue to form a straight line, much less a straight line that is perpendicular to a given line? In fact, the special theory of relativity says very explicitly that two observers in motion relative to each other, each watching the line move as described by ibn al-Haytham, will *probably not* agree that it remains perpendicular to the given fixed line. The difficulty, it turns out, concerns the notions of simultaneity and temporal order of events. When the base point moving along the line is at point Q, where is the rest of the moving line? The presence of each of these points at one place or another is a different event, and two observers will not generally agree as to which of those events is simultaneous with the passage of the moving line through point Q.

The reconsideration of the notion of simultaneity required by special relativity reinforces Omar Khayyám's point. The event that is occurring at a given location

[1]Until the advent of projective geometry, all geometers thought of a line as what we call a closed line segment. Euclid says that the extremities of a line are points. We shall use the word *line* in this traditional sense in order to save writing the word *segment* excessively many times.

at a given time (the presence or absence of a given particle, for example) depends on who is observing. Perpendicularity of two lines requires that a certain configuration among three variable points remain constant at all times. One of the points, say Q, is the moving point of intersection of the two lines; the second point, say P, is fixed on the moving line; and the third, say R, lies on the fixed line. As we are about to discuss, when special relativity is admitted, it is impossible to state in an observer-independent way what the relative positions of three particles are at a given instant of time. It follows from the relativistic equations of transformation between observers that the temporal order of two events occurring in different locations may depend on the observer, and it is easy to see that lines regarded as mutually perpendicular by one observer are normally not mutually perpendicular as measured by a second observer. In terms of spatial coordinates, this disagreement comes about because of the *FitzGerald–Lorentz contraction*,[2] whereby the length one observer ascribes to a line in the direction of the other observer is found to be shortened by the factor $\sqrt{1-u^2/c^2}$ when measured by the other. Here u is the mutual speed with which the two are moving relative to each other, and c is the speed of light, approximately 300,000 km/sec. This inconsistency of spatial measurement, in turn, arises from the fact that the observers disagree as to which of two events occurred earlier: They are unable to agree as to where one end of the line is in relation to the other at a given instant of time, since simultaneity of two events occurring in different places is not observer-independent. The details of this phenomenon will be explored below. We are now going to flesh out these abstract ideas with a scenario taken from the modern world.

1.1. The car wash puzzle. Imagine a very long limousine parked in a car wash, the limousine being exactly as long as the car wash. It would just fit inside, and the attendant would be able to close the doors at both ends of the car wash. Now imagine that same limousine driving through the car wash (at any speed you like, but imagine it to be a very high speed). Is there an instant of time when the limousine is entirely inside the car wash? The limousine driver will say no: Due to the FitzGerald–Lorentz contraction, the car wash shrank in length, and the limousine wouldn't fit inside. The car wash attendant will say yes: Due to the FitzGerald–Lorentz contraction, the limousine shrank in length and fitted inside with room to spare. Who is right here?

The explanation involves the observer-dependence of the concept of simultaneity. We are looking at two events here that take place in different locations. One event is the rear end of the limousine entering the car wash. The other is the front end of the limousine leaving the car wash. Those events occur in the opposite order for the two observers. While they share the same four-dimensional space-time and agree about the four-dimensional "proper-time" interval between the two events—taking one of the events to have occurred at time zero and at the origin of the spatial coordinates in both systems, while the other occurred at time t and at a point (x, y, z), that interval is $t^2 - (x^2 + y^2 + z^2)/c^2$ and is the same for both of them—they do not agree about the *scales* on either the line in space along the common direction of motion or on the time axis. In other words, t and at least one of the spatial coordinates, say x, are different for the two observers even though the four-dimensional interval is the same for both of them. The answer to the question

[2]Named after George Francis FitzGerald (1851–1901) and Hendrik Antoon Lorentz (1853–1928).

is that the question does not make sense. The individual time order of two events A and B will be the same for all observers only if a ray of light could set out from the location of one of them, say A, at the time A occurred and arrive at the location of event B before B occurs there. That is to say, the four-dimensional interval between A and B is positive—"timelike," as physicists refer to it. Because this four-dimensional interval is the same for all observers, any two will agree whether this is the case or not.

Another way of saying that the interval between events A and B is timelike is to say that an observer could physically be present at both events A and B. The proper time interval[3] between the two events (for any observer) is obtained by parameterizing a path from A to B, say $r \mapsto \bigl(t(r), x(r), y(r), z(r)\bigr)$, where A corresponds to r_0 and B to r_1, and then integrating the infinitesimal proper time interval ds, which is given by

$$ds^2 = dt^2 - \frac{1}{c^2}\,dx^2 - \frac{1}{c^2}\,dy^2 - \frac{1}{c^2}\,dz^2\,.$$

Thus we find that the proper time interval is

$$\Delta s = \int ds = \int_{r_0}^{r_1} \sqrt{\bigl(t'(r)\bigr)^2 - \Bigl(\frac{x'(r)}{c}\Bigr)^2 - \Bigl(\frac{y'(r)}{c}\Bigr)^2 - \Bigl(\frac{z'(r)}{c}\Bigr)^2}\,dr\,.$$

If an observer is present at all the events $\bigl(t(r), x(r), y(r), z(r)\bigr)$, then in that observer's coordinate system, $x(r) = y(r) = z(r) = 0$ at time $t(r)$ for all r. Hence $x'(r) = y'(r) = z'(r) = 0$ for that observer, and (assuming B is later than A for that observer), $t'(r) \geq 0$. Therefore the time interval recorded by that observer is

$$\int_{r_0}^{r_1} \sqrt{\bigl(t'(r)\bigr)^2}\,dr = \int_{r_0}^{r_1} t'(r)\,dr = t(r_1) - t(r_0)\,.$$

In short, *the proper time interval between two events is the time interval recorded by an observer who was present at both.* More generally, if an observer assigns the same spatial coordinates to two events, then the proper time interval between them is just the time difference recorded by that observer.

If there were any observer who perceived B as occurring before A, then paradoxes might result if the interval is timelike—that is, a ray of light could leave the location of A at the time A occurred and arrive at the location of B before B occurred (which is equivalent to saying that $\Delta s^2 > 0$). A second observer at the site of event B could transmit information about event A before event B occurs and thus the observer who perceives B as occurring first could get historical information about event A *before it occurred.* That observer would be "remembering the future."

If the space-time interval between A and B is not positive—it is "spacelike," as physicists say—there is no observer-independent temporal ordering between two events occurring in different locations, and there is no absolute sense in which one

[3]The terminology is due to Hermann Minkowski (1864–1909), who gave it this name in a paper [**60**] published in the last year of his life. Minkowski had read this paper at a meeting in Köln on 21 September 1908. He actually gave the name "proper time" (*Eigenzeit*) to a quantity that has the physical dimension of length, namely $-dx^2 - dy^2 - dz^2 - ds^2$, where s is time made into an *imaginary length* by means of what he called (p. 86), the "mystical formula" $3\cdot 10^5\ \mathrm{km} = \sqrt{-1}\,\mathrm{sec}$. The concept of proper time had been introduced earlier, however, by Lorentz, who called it *local time.*

event occurred before the other. A geometric explanation of this puzzle can be found by working Problem 1.19 below.

The car wash puzzle throws some doubt on ibn al-Haytham's intuition. The points P, Q, and R are in different locations, and different observers will not agree as to their relative configuration at a given time. For two observers O and O' in relative motion, both watching the two lines slide along each other, there will not in general be any agreement as to the angle between the two lines. Unless the fixed line is the line joining O and O', the two will not both agree that the moving line is perpendicular to it. In that respect, Omar Khayyám's objection gains force when special relativity enters the picture. Nevertheless, it remains true that both observers are carrying out their measurements using Euclidean principles, and they agree that the end of the moving line *not* on the fixed line is describing a line parallel to the fixed line. Thus, one can retain ibn al-Haytham's conclusion while rejecting the considerations that led him to it.

In the following sections, we shall investigate the relations among the speeds and directions of three or more observers, assuming that the space-time coordinates of any pair are reconciled using the equations of special relativity. As we shall learn, there is an inherent difficulty in regarding the velocity spaces used by two different observers as having any vectors in common, even though the equations of transformation from the coordinates of one to those of the other assume that the vector spaces representing locations (not velocities) do share a line, namely the line joining the origins. A unifying theory, which unfortunately lies beyond the scope of the present book, is the *Lorentz group*, the six-dimensional Lie group of transformations of $\mathbb{R}^4$ that preserve the four-dimensional interval between events.

2. Synchronization in Newtonian Mechanics

I. Absolute, true, and mathematical time, of itself, and from its own nature, flows equably without relation to anything external, and by another name is called duration: relative, apparent, and common time, is some sensible and external (whether accurate or unequable) measure of duration by the means of motion, which is commonly used instead of true time; such as an hour, a day, a month, a year.

II. Absolute space, in its own nature, without relation to anything external, remains always similar and immovable. Relative space is some movable dimension or measure of the absolute spaces; which our senses determine by its position to bodies; and which is commonly taken for an immovable space; such is the dimension of a subterraneous, an aerial, or celestial space, determined by its position in respect of the Earth. Absolute and relative space are the same in figure and magnitude; but they do not remain always numerically the same. For if the Earth, for instance, moves, a space of our air, which relatively and in the respect of the Earth remains always the same, will at one time be another part of the same, and so, absolutely understood, it will be continually changed.

Isaac Newton, *Mathematical Principles of Natural Philosophy*, Scholium to Definitions I–VIII ([**63**], pp. 8–9).

To the ancient Greeks, most of whom adhered to a geocentric cosmology, there was a perfect geometric representation of the universe: It was a series of concentric spheres, whose center was at the center of the Earth. The absolute geometric objects corresponded perfectly to what they believed was the nature of physical space. Later, when heliocentric astronomy replaced the geocentric model, the sphere was banished as the embodiment of physical space and was replaced by the unbounded three-dimensional flat Euclidean space of solid geometry, soon algebraized by the use of Cartesian coordinates.

In Newtonian mechanics, all observers share this common three-dimensional Euclidean space and can agree about the scale on a time axis independent of space. The location and the time of an event are separate issues, and all observers agree about both of them. Underlying this view, as expressed by Newton in the quotations above, is an unconscious metaphysical idea, an intuitive notion that there is a well-defined "now" that is common to all locations in the universe. Newton referred to it as "absolute time," that is, the pure article, undiluted by any of the observable "events" such as the hands on a clock assuming a certain position by which we actually determine the time. That inferior, diluted form of time is what Newton called "relative time." Many modern philosophers hold the view—one congenial to the present author—that what Newton called absolute time is a pure mental creation and is unobservable and unknowable, analogous to the Aristotelian distinction between "substance" and "accidents." These philosophers hold that we can talk meaningfully only about Newton's relative time (or, in the Aristotelian context, only about the accidents). While these philosophers have very likely adopted an unassailable position, they may be underestimating the value of human imagination in the creation of physical theories. Other concepts of mathematical physics—force, for example, which Bertrand Russell called "a convenient fiction"—can be similarly challenged by an obstinately practical-minded philosopher. Yet, as a guide to thought, leading to insights about the properties of observable objects, the ability to hypostatize[4] abstract concepts, picturing them as if they were observable physical objects, can be of enormous value, provided the hypothesis of their reality leads to the prediction of the result of a physical measurement. The concept of gravity is a good example of such a mental creation.[5] It may be only a convenient fiction, but it is a *very* convenient one.

Newton's attempt to define absolute space without reference to any observable objects is subject to the same metaphysical objections as his definition of time. Mathematicians and physicists find it useful to picture space as endowed with imaginary "mileposts" labeling each of its points with three numbers giving the distances of that point in front of, to the right of, and above, a fixed imaginary reference point (the origin). That is absolute space, and it "exists" only as a mental construct. For the mathematician, the space $\mathbb{R}^3$ *actually is* just the set of ordered triples of real numbers (x, y, z). This is "absolute Euclidean space." Such a picture

[4]From the Merriam–Webster on-line dictionary. "*hypostatize:* To attribute real identity to (a concept)." The word comes from the Greek word *hypostatos* (ὑπόστατος), which has the same two roots as the Latin-derived word *substance*.

[5]Originally, *gravity* (*gravitas*) was just the Latin word for heaviness. It was a property possessed by such things as earth, air, and water, but not by fire, which had the opposite property of *levity* (*levitas*) or lightness. The mental picture of gravity as a *force*, a *thing* "existing" independently of bodies having the *property* of gravity, even though no one has any clear ideas as to *what* that thing is, remains useful.

cannot be used experimentally, however. As Newton suggested, we need *observable physical* objects in order to define a three-dimensional frame of reference in which the locations of all objects can be given in terms of three coordinates. He said that there might be, in the remote depths of interstellar space, a body perfectly at rest, but that it is impossible for us to infer from the positions of the bodies we observe which, if any, of them is at rest relative to absolute space. Nevertheless, he thought people writing philosophical treatises such as his *Principia* ought to "abstract from our senses, and consider things themselves, distinct from what are only sensible measures of them." In such statements, he sounds very much like Plato, who argued (in Book 7 of the *Republic*) that astronomy becomes useful only after one makes it a subject to be perceived by the mind, ignoring what the senses perceive in the heavens.

Although it was eventually forced to yield to hard stubborn facts, such as the absence of any detectable medium that conducted light waves, the mental habit of hypostatizing absolute time and space is still with us, and has a great deal of appeal for minds that evolved in a world of slow velocities and distances that lie within the range of perception of human senses, things that we can grasp intuitively. One feels, for example, that there really is a difference between saying the Sun orbits the Earth once a day and saying that the Earth rotates on its axis once a day. Newton himself (still in the section of definitions at the beginning of the *Principia*) argued that we could distinguish absolute motion from relative motion by looking at the case of circular motion about an axis. He pointed out that if a bucket is hung on a twisted cord and the cord allowed to unwind, the bucket spins, eventually causes the water to spin, and then the water rises up the sides of the bucket. Surely that suggests some absolute frame, at least for rotational motion, since the relative motion of the bucket and the water was greatest at the beginning, when there was no tendency for the water to rise; when the relative motion was least (the water had attained its maximal rotational velocity, equal to that of the bucket), the tendency to rise was at a maximum. That picture is a familiar one, and leads naturally to the idea that there is an absolute set of invisible axes relative to which rotation occurs. Such a picture is at the basis of a scene from C. S. Lewis' 1944 science-fiction novel *Perelandra*.

> *It was born in upon him that the creatures were really moving, though not moving in relation to him. This planet which inevitably seemed to him while he was in it an unmoving world*—the *world, in fact—was to them a thing moving through the heavens. In relation to their own celestial frame of reference they were rushing forward to keep abreast of the mountain valley. Had they stood still, they would have flashed past him too quickly for him to see, doubly dropped behind by the planet's spin on its own axis and by its onward march around the Sun.* [*Perelandra*, Chapter 16.]

The idea of absolute time and space still lingers, and even general relativity appears at first sight to require some replacement for it.[6] Beyond doubt, a great deal has been achieved with this concept through Newtonian mechanics. By the twilight years of this concept, at the end of the nineteenth century, it had

[6]One such replacement is known as *Mach's principle*, after Ernst Mach (1838–1916). We shall return to that discussion in Chapter 8.

been brilliantly supplemented by vector analysis, an outgrowth of William Rowan Hamilton's quaternions, leading to an elegant, compact mathematical notation for physical laws that could easily be turned into a set of coordinates at any time when numerical results were needed. In this cozy, familiar-seeming picture of the universe, two observers[7] O and O' in uniform relative motion have little difficulty reconciling their observations. It is not difficult to convert the coordinates that O uses for the time and place of an event with those used by O', whether the event is something that occurs in daily life, or a celestial phenomenon. When the conversion of the time and space coordinates of an event cannot be separated into a conversion of the time coordinate independent of the conversion of the space coordinates, however, as we now know to be necessary, it becomes more difficult for O' to translate the coordinates used by O into his own coordinates. Let us now make a more detailed investigation of this problem.

2.1. The speed of information and Newtonian synchronization. To picture the Newtonian scheme in a more concrete manner, let us imagine a pair of mid-eighteenth-century twins. To make the backstory more colorful, let's assume that they are English and are named John and Mary. As a further part of the backstory, let us assume that Mary has emigrated to Massachusetts with her husband, one James Foster, while her brother John has remained in London. To keep up with each other's lives, each sends a local newspaper to the other along with a letter every week.

In that case, perhaps in the winter of 1763, when both of them would have been eager for news of the end of the Seven Years' War in Europe (known as the French and Indian War in the United States), Mary might receive the *London Chronicle or Universal Evening Post*, Vol. 13, No. 944, and on p. 33 (Saturday, 8 January to Tuesday, 11 January 1763), read the following story:

> *On Saturday the river Thames was frozen so hard at Isleworth that a fair was kept on it all day. A large booth was erected in which was sold beer and other liquors, and in which a leg of mutton was boiled for the company. There was a round-about for children to ride in and all sorts of toys sold, as at other fairs. Great numbers of people came from the adjacent parts to see it.*

And perhaps John would, about the same time, read the following sad note—confirmed by an accompanying letter from Mary—on p. 3 of the *Boston Evening Post* of January 10, 1763:

> *Last Saturday died at Dorchester, after a few Days Illness, James Foster, Esq; of that Town.*

We have made it easy for Mary and John to synchronize their time-keeping, which in their case involved calendars rather than clocks. Since the two would have been using a common calendar—the Gregorian, recently (1752) adopted throughout the British Empire—the reference to Saturday in both stories would cause both

[7]For convenience, two observers are just two people in motion relative to each other, and to avoid clumsy he/she constructions, we shall simply assume they are both male. Each is pictured as located at the origin of a set of three spatial coordinates and carrying a clock. Reconciling the spatial coordinates and the clocks between the two observers is the central problem of the present chapter.

siblings to conclude that these two events occurred simultaneously. (We shall ignore the five-hour difference in solar time between London and Boston and assume that events that occurred on a given date in Massachusetts also occurred on that same date by London reckoning, even though this is often not the case.) That is, Mary would conclude that the fair her brother attended on the frozen Thames was going on at the same time that her husband was on his deathbed. And John would agree that, unaware of this sad event, he had been enjoying beer and leg of mutton while his brother-in-law was breathing his last. The two events were simultaneous; and even though neither sibling observed both of them, they would have been able to reconcile their timekeeping by using a common calendar.

Even without that calendar, however, each of them could have computed the time of the event at which the other was present, provided they knew (1) the time t_0 elapsed between those events and the dispatch of the report aboard the ship, (2) the average speed v_0 at which the ship sailed, and (3) the distance d_0 by sea from London to Boston.[8] Dividing the distance d_0 by the speed v_0 would yield the time that the newspaper was en route. Subtracting the sum t_0+d_0/v_0 from the time that the news arrived would give the time of the event on the *common* calendar both were using. To make the analogy with physics, the point of Newtonian mechanics is that there *is* a universal method of measuring time, usable by everyone, and it is not affected by any state of motion of any observer. Likewise, there is a universal method of specifying locations through a system of three rectangular coordinates (x, y, z), and everyone can use this system, thereby always agreeing on the absolute location of any particle at any absolute time.

This small anecdote—James Foster was a real person who died in Dorchester on 8 January 1763, but his having emigrated from London with a wife named Mary is fiction—illustrates two points of importance to physics. First, information travels at a finite speed.[9] Second, if an event in one place is the cause of a second event in another place, it must be possible for information about the first event to reach the second place before the second event occurs: John cannot write to his sister to express condolences and perhaps suggest that she consider returning to London until *after* he receives news of his brother-in-law's death. Thus, travel at eighteenth-century speeds, with the eighteenth-century "speed of information"[10] being what it was, has rather weak effects on the lives of ordinary people. If, anachronistically, they both had accurate calendar-watches that they synchronized when Mary departed from London, they would find when she returned that their watches were still synchronized, as far as they could tell. In particular they would find that they had both aged by the same amount, and they would agree as to the total distance Mary had traveled. We shall bring these siblings back to the stage later in the present chapter, moving their stories forward in time by some three centuries and making Mary's travel a high-speed journey to a planet orbiting

[8]As John and Mary were citizens of the British Empire in the eighteenth century, that distance would have been given as roughly 3300 miles (say 5300 kilometers).

[9]The claims that paranormalists make on behalf of precognition and clairvoyance do not hold up well under scrutiny.

[10]We shall make formal use of the concept of speed of information starting in Chapter 4, reserving the symbol v_0 for the arbitrary hypothetical value we assign to it in the Newtonian world. In Newtonian theory, it could have any positive value, but of course had a rather small one in the eighteenth century. Once communication by radio was invented, v_0 became coincident with the speed of light, for which we use the symbol c. Further increases in its value are not expected.

a nearby star. At that time, as we shall see, travel has rather more noticeable—relativistic—effects on a person, and their maps and timetables for the journey will not agree. For their final bow, in Chapter 7, we'll move them ahead by yet another three centuries and ask Mary to travel to a black hole. The discrepancies in their measurements of time and space will then be still more noticeable.

Physics is much more abstract than everyday life. The kind of events reported in news media are described by the "five W's": what, who, where, when, and why. That is, they tell what happened (picnic on the frozen Thames, death of James Foster), who was involved (large numbers of people, James Foster), when (8 January 1763), and (in reporting complicated issues) a context that makes the event comprehensible to the reader or viewer. We have confidence in the accuracy of reports if all journalists agree on these five points. Physics, in contrast, pays no attention to individual people and their motives. Accordingly, in physics we have observers, not journalists, and they are concerned only with what happened, where, and when. Now if two observers are to reconcile their observations at all, they surely have to agree as to what happened. (Say, one particle collided with another.) In Newtonian physics, two observers can reconcile their observations so as to agree separately about the time and location of the event. In special relativity, in contrast, they cannot. Instead of having a where and a when to reconcile, each observer has a "where-when," and one where-when is reconciled with another through the Lorentz transformation that will be introduced below.

From the concept of events, we get the notion of a space-time in which an event is simply a set of four coordinates $(t; x, y, z)$, t being the absolute Newtonian time of the event, and (x, y, z) its spatial location in terms of a conventional three-dimensional coordinate system. In one sense, it can be argued that the speed of information in Newtonian physics is infinite. In such equations as Laplace's equation, the heat (diffusion) equation, and Newton's law of gravity, any perturbation of the controlling initial/boundary conditions is propagated instantly to the solution at all points of space for all later times. The only exception is the wave equation, in which disturbances of the initial condition propagate at a finite speed. This unrealistic feature of the Newtonian equations is one reason for preferring the relativistic ones. The one exception—the wave equation—which governs electromagnetic radiation, lies at the very heart of the special theory of relativity.

It is hoped that the reader finds none of this subsection difficult to understand. Most readers will, more likely, be impatient at being patronized by such a detailed discussion of what is, after all, only common sense. The trouble is that Common Sense is Newtonian, but the physical universe isn't. We have included this subsection in order to lay down a detailed background of concepts that can be modified in intuitively reasonable ways, one step at a time, all the way to the general theory of relativity.

We are now in the realm of abstract mathematics, and to tie it to the physical world, we need an interpretation of the mathematical object $(t; x, y, z)$. In the Newtonian scheme, we can think of (x, y, z) as the location of an identifiable particle at time t, perhaps a proton or electron. Although these physical bodies do occupy some volume, we can idealize them as having all three of their geometric dimensions equal to zero. If the particle moves, we can identify its position at time t with a vector $\boldsymbol{r}(t) = \bigl(x(t), y(t), z(t)\bigr)$. Along with that position, the particle has other numbers associated with it, such as its mass m, possibly its electrical charge q, its velocity $\boldsymbol{r}'(t)$, its acceleration $\boldsymbol{r}''(t)$, its momentum $m\boldsymbol{r}'(t)$, the force acting on it $m\boldsymbol{r}''(t)$, its kinetic energy $(1/2)m|\boldsymbol{r}'(t)|^2$, and, if the forces that are acting on it are conservative, its potential energy $V\bigl(\boldsymbol{r}(t)\bigr)$, which depends on its location. But everything is defined in terms of the mass m and the four coordinates $(t; x, y, z)$, with t having the physical dimension of time, and the other three having the physical dimension of length. We are now going to explore the concept of ordered time in Newtonian mechanics from this more abstract point of view.

2.2. Four kinds of time. In the story told above, we see two related events occurring. One is a primary event, but secondary to it is another event, namely the *observation* of the primary event. The second event occurs later because information requires time to travel from the location of the primary event to the location of the secondary event. To "abstractify" these considerations and fit them into mechanics, yet at the same time present a simple model for understanding, we find it useful to consider a moving particle that is being viewed by an observer. For the time being, let us assume that the observer is using an orthogonal system in which the Pythagorean theorem holds, and we think of the observer as sitting at the origin of this coordinate system. Since Newtonian time is absolute, we can assume that the all clocks used by all observers are properly synchronized with one another. Finally, we shall assume that information travels at some fixed speed v_0 in any direction. If, at time t an observer at rest at the origin of Newton's universal space receives information that something interesting is happening to a particle at time t that is located at the point with universal coordinates $\boldsymbol{r} = (x, y, z)$ at time t, then that observer, taking account of the finite speed of information, will conclude that that event *really* happened at an earlier time s given by

$$s = t - \sqrt{x^2 + y^2 + z^2}/v_0 = t - |\boldsymbol{r}|/v_0\,.$$

The time s is the *Newtonian universal time* showing on a clock at the point (x, y, z) when the event occurred. It will be the same for any observer viewing the event. The *observation* time t, which differs from one observer to another, is the value that universal time has when information about the event reaches the observer. For two events that occur at locations $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ and are observed at the origin at times t_1 and t_2 respectively, the time interval between the events themselves is

$$\Delta s = s_2 - s_1 = \Bigl(t_2 - \frac{|\boldsymbol{r}_2|}{v_0}\Bigr) - \Bigl(t_1 - \frac{|\boldsymbol{r}_1|}{v_0}\Bigr) = \Delta t - \frac{\Delta|\boldsymbol{r}|}{v_0}\,.$$

This expression will be the same for any two observers viewing the events, even though the information about the events will reach them at different times and the coordinates they assign to the locations of the events will generally be different. In order to think clearly about the four-dimensional relativistic world whose points are "events" $(t; x, y, z)$, we need to keep this "speed of information" v_0 in the background. In relativity, it will be the speed of light in free space, but in Newtonian mechanics, it can be any convenient positive speed. In Newtonian mechanics, the difference between observation time and universal time is merely the time required for information to travel from the site of the event to the observer. In relativity, that Newtonian adjustment is taken for granted as having been made by any given observer O, who therefore has a clear concept of simultaneity throughout his own personal Euclidean space. But communication with another observer O' in motion relative to O is complicated, since neither the similar synchronization O' has carried out nor his spatial measurements agree with those of O. The discrepancy all hinges on one issue, which is precisely time-keeping. Since we no longer have a universal time, we shall distinguish between Newtonian and relativistic measurements by calling the observer's time *laboratory time*. The best replacement we can get to help our two observers reconcile the order of events is yet another kind of time we shall call *proper time*, which we use as the replacement for universal time.

2.3. Proper time. The distinction we have just made between observation time and universal time in Newtonian mechanics becomes much more important in relativity, where the interval between the time s shown by a clock attached to a moving particle and the time t recorded by someone observing the particle is given by "enlarging" the Pythagorean theorem so as to include a time dimension. On the infinitesimal level, as mentioned above,

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

Here, c is the speed of light, the fastest speed at which information can be transmitted. The analogy is not perfect, as will be seen, since in relativity it is *impossible* for two observers in relative motion, each having an accurate clock, to synchronize those clocks. In relativistic mechanics, we need to imagine *two* clocks associated with any moving object that is being observed. The time t that the observer records, which we just agreed to call *laboratory time*, comes from a clock synchronized with the observer's clock and attached at the point of the observer's space that the moving object is passing through at a given instant; it is, from the point of view of someone riding on the moving object, not keeping accurate time. The proper time s is what is shown on a clock attached to the object. Although the first clock is synchronized with his own local clock, the observer will still have to make the correction for the time it takes for a signal from that clock to reach him in order to assign a time to this event. But even after that correction is made, the time the observer records for an event will not agree with the time broadcast at the same instant by the second clock, attached to the moving object.

All that will be explained in detail below. In the relativistic equations relating the space-time coordinates of two observers, the correction for time lag is already taken into account in the variable t. *The laboratory time t of an event is not the time at which the observation was recorded (what we just called "observation time"), but rather the time at which the event actually happened in the observer's personal space-time, taking the Newtonian adjustment into account.* The Newtonian correction amounts to taking account of the *time delay* involved in transmitting information about the event, and any two observers who have made that correction will agree on the time of the event. In the relativistic model, however, even after correcting for the time lag due to a finite speed of information, different observers will not in general ascribe the same space-time coordinates to an event.

There is in fact only the one invariant across all these coordinate systems, namely the infinitesimal squared-increment of proper time ds^2, which is expressed as a quadratic form in the differentials of the four coordinates. It is the same for any two observers in relative motion along a straight line at a constant speed. All of this falls into the domain of the special theory of relativity, which will be discussed in the next chapter. In this theory, two observers using different coordinate systems can "talk to" each other, if their coordinates are related by a *Lorentz* transformation, which is a linear transformation of $\mathbb{R}^4$ that preserves the quadratic form ds^2 just written.

2.4. Homogeneous coordinates*. The use of $\mathbb{R}^4$ to represent the events in the Newtonian universe has the seeming disadvantage that the physical dimensions of the four coordinates are different. Really, what we have is a three-dimensional space of distances and a one-dimensional space of time, and they are completely independent of each other. Nevertheless, we would like to be able to talk about the *interval* separating two events $(t_1; x_1, y_1, z_1)$ and $(t_2; x_2, y_2, z_2)$. That will involve somehow adding the time interval to the spatial interval so as to get a measure of the full interval, as we just did in the previous subsection. It is a well-known principle of all applied mathematics that concrete numbers of different types cannot be added. The way to solve this problem is to assign some equivalence to time and distance. We used the "speed of information" v_0 for that purpose in the example above.

We do this constantly in everyday speech, reporting distances as if they were times. Thus we say that it is "a twenty-minute walk" from one's home to the grocery store, or that a certain city is "two hours away by car." In military language, a scout at one time might have reported that the enemy camp is "three days' march from here." And astronomers tell us that the nearest star is approximately "four light-years" away. In all these cases, we specify two independent quantities: (1) a unit of time (minutes, hours, days, years) and (2) a speed (walking speed, average driving speed, marching speed, the speed of light) to convert distance into time. The important fact to be kept in mind is that in Newtonian mechanics, *both* of these standards are pure conventions, and any unit of time or speed will do. Thus there are *two mutually independent* arbitrary units in Newtonian mechanics. Once they are chosen, actual times and distances may (theoretically) be represented by any real numbers.

One consequence of this Newtonian independence of space and time is that any two events occur in a definite temporal order, one that will be agreed upon by any two observers. To reach agreement, each of them has only to compute the universal time of the two events, which will be the same for both observers, provided they synchronize their clocks at any given time. Then Event 1, occurring at universal time s_1 precedes Event 2, occurring at universal time s_2 if $s_1 < s_2$, and any two observers will agree whether this is the case or not. (In practice, the two observers may not be able to measure time precisely enough to say which of two events was earlier, but the theoretical ordering remains.) This intuitive ordering, as we saw in the car wash puzzle, gets shattered by the special theory of relativity, in which the ordering assigned to two events by one observer may be the opposite of the order assigned by another observer.

2.5. The Galilean transformation. Before introducing the special theory of relativity formally, we give the coordinate transformations in Newtonian mechanics linking two observers that are both using the same x-axis and have clocks that are synchronized. If Observer O_2 is moving with speed u along the x-axis shared with Observer O_1 and the axes of the two systems coincide at time $t_1 = t_2 = 0$, then the coordinates the two will assign to an event are related by

$$\begin{aligned} t_2 &= t_1 \,, \\ x_2 &= x_1 - ut_1 \,, \\ y_2 &= y_1 \,, \\ z_2 &= z_1 \,. \end{aligned}$$

This transformation is called the *Galilean* transformation, after the pioneering scientist Galileo Galilei (1564–1642).

REMARK 1.1. Newtonian space is Euclidean and each point in it, regarded as a vector $\boldsymbol{\xi} = (x, y, z)$, has a squared-distance from the origin given as its dot product with itself:

$$|\boldsymbol{\xi}|^2 = \boldsymbol{\xi} \cdot \boldsymbol{\xi} = x^2 + y^2 + z^2 \,.$$

The vector notation for the dot product (invented in the late nineteenth century) is very useful because of the compact expression it gives to many physically important quantities. The Euclidean structure of space singles out certain coordinate systems as "preferred," namely those that are *orthonormal*, meaning that the basis coordinates $\boldsymbol{\xi}_1 = (1, 0, 0)$, $\boldsymbol{\xi}_2 = (0, 1, 0)$ and $\boldsymbol{\xi}_3 = (0, 0, 1)$ satisfy $\boldsymbol{\xi}_i \cdot \boldsymbol{\xi}_j = 0$ if $i \neq j$ and $\boldsymbol{\xi}_i \cdot \boldsymbol{\xi}_i = 1$. If we confine ourselves to orthonormal coordinate systems, it does not matter which particular one we choose, since the square-distance is given by the same expression: $|\boldsymbol{\xi}|^2 = x^2 + y^2 + z^2$ whenever $\boldsymbol{\xi} = x\boldsymbol{\xi}_1 + y\boldsymbol{\xi}_2 + z\boldsymbol{\xi}_3$. That is one huge advantage of orthogonal systems: The dot product is invariant under orthogonal transformations, which are defined by that property: $T\boldsymbol{\xi} \cdot T\boldsymbol{\eta} = \boldsymbol{\xi} \cdot \boldsymbol{\eta}$.[11]

This seems an appropriate point to foreshadow certain other aspects of physics that are affected by the use of alternative systems of coordinates. We have in mind particularly the concept of kinetic energy. Assume that O_1 and O_2 are both fixed at the same origin ($u = 0$), and O_2 continues to use an orthonormal coordinate system but that O_1 is using a general coordinate system. For what we want to do, we need to be slightly more formal and systematic about our labeling. Henceforth, we let $x_i = x_i^1$, $y_i = x_i^2$ and $z_i = x_i^3$, $i = 1, 2$. Then for some constants g_{ij}, $i, j = 1, 2, 3$, we have

$$\begin{aligned} x_2^1 &= g_{11}x_1^1 + g_{12}x_1^2 + g_{13}x_1^3 \,, \\ x_2^2 &= g_{21}x_1^1 + g_{22}x_1^2 + g_{23}x_1^3 \,, \\ x_2^3 &= g_{31}x_1^1 + g_{32}x_1^2 + g_{33}x_1^3 \,. \end{aligned}$$

[11] The cross product $\boldsymbol{u} \times \boldsymbol{v}$ is invariant under a rotation, that is, an orthogonal transformation whose determinant is 1. For that reason, physicists sometimes refer to the cross product as a *pseudo-vector* reserving the term *vector* for vectors that are invariant under all orthogonal transformations.

Then the kinetic energy T of a particle of mass m that O_2 observes to be at position $\boldsymbol{r}_2(t)$ at time t, is

$$\begin{aligned} T = \frac{1}{2}m|\boldsymbol{r}_2'(t)|^2 &= \frac{1}{2}m\big((x_2^{1\prime})^2 + (x_2^{2\prime})^2 + (x_2^{3\prime})^2\big) \\ &= \frac{1}{2}m\sum_{j=1}^{3}\big(g_{j1}x_1^{1\prime} + g_{j2}x_1^{2\prime} + g_{j3}x_1^{3\prime}\big)^2 \\ &= \sum_{j=1}^{3}\sum_{i=1}^{3} t_{ij}(x_1^{i\prime})(x_1^{j\prime})\,, \end{aligned}$$

where

$$t_{ij} = m\sum_{k=1}^{3} g_{ki}g_{kj}$$

if $i \neq j$, and

$$t_{ii} = \frac{1}{2}m\sum_{k=1}^{3} g_{ki}^2\,.$$

In Newtonian mechanics, it is perfectly legitimate—though one might think it foolish—to use general coordinates. Still, one might be studying the crystalline structure of a body and prefer axes that follow the lines of symmetry of the crystals. In that case, we might actually use oblique coordinates. If we do so, we see that we need more than the simple scalar equation $T = (1/2)m|\boldsymbol{r}'(t)|^2$ to keep track of kinetic energy. In its place we need what is called a *tensor*, which in this case is a bilinear mapping of pairs of velocity vectors $\boldsymbol{u} = (u^1, u^2, u^3)$ and $\boldsymbol{v} = (v^1, v^2, v^3)$:

$$T(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i,j=1}^{3} t_{ij}u^i v^j\,.$$

This same tensor, with the factor of $m/2$ divided out of every entry, gives the square of the infinitesimal element of arc length $ds = \sqrt{dx_1^2 + dy_1^2 + dz_1^2}$ in coordinates $x = x_2^1$, $y = x_2^2$, $z = x_2^3$:

$$ds^2 = \sum_{i,j=1}^{3} g_{ij}\,dx_2^i\,dx_2^j\,,$$

where $g_{ij} = 2t_{ij}/m$. This last tensor is of basic importance throughout differential geometry. It was the starting point for modern abstract differential geometry, introduced by Bernhard Riemann (1826–1866) in his 1854 inaugural lecture. It soon came to be called the *fundamental tensor* by Einstein and others. We shall call it the *metric tensor*, since it gives the metric by which intervals are measured on an abstract manifold. Here we have our first hint of the intimate connection between differential geometry and mechanics: the tensor that defines the geometry of a manifold also serves to convert velocities into kinetic energy, through exactly the same bilinear operation on a pair of vectors in both cases. As we progress from Newtonian mechanics to general relativity, that metric-energy connection will serve as a guide. We will think of the metric coefficients (the g_{ij}) as potential energy functions. Notice that all of this insight comes about because we attempted to free ourselves from dependence on particular coordinate systems. As long as we confined ourselves to orthonormal coordinates, where the metric is $ds^2 = dx^2+dy^2+dz^2$ and the kinetic energy is $m|\boldsymbol{r}'|^2/2$, we wouldn't necessarily think in terms of a matrix.

While we would not normally bother with this tensor in Newtonian mechanics, the analogous four-dimensional concept in relativity will turn out to be very useful to us. We shall see this at the end of Chapter 2 and again in Chapters 6 and 7. Certain bilinear mappings of pairs of velocity vectors occur quite naturally in the equations of geodesics, which lie at the heart of relativity theory. By putting this example here, we are foreshadowing some important results that will come later and preparing the reader to adjust to a new way of thinking about mechanics.

REMARK 1.2. The reader may also be wondering why we chose to measure distance as time, converting it via a conventional standard velocity v_0 (taken to be c, the speed of light, in relativity). Why not instead convert time to distance by defining $\tilde{t} = v_0 t$? After all, we generally find it easy to measure distance. We can, in the simplest case, carry a ruler around with us calibrated in standard units of length, such as millimeters, and determine the distance between two nearby points. Time, on the other hand, is a rather mysterious, mystical thing (see the discussion in Chapter 8). In order to measure it, we have to select some process that we accept as proceeding at a uniform rate—the dripping of water through a hole, or sand in an hour-glass, or the swing of a pendulum, or the unwinding of a watch spring, or the vibration of the crystal in a digital watch or the right ascension of a star—and use that process as a measure of time. It is not intuitively obvious that all these ways of measuring time are even mutually consistent. The now old-fashioned clock with hour and minute hands goes in exactly the opposite direction from the conversion we made, measuring time by the lengths of the arcs traversed by the tips of the two rotating hands. And we have all learned geometry by thinking of lines as lengths. Why this nonintuitive, seemingly needless complication? It is certainly possible to express time as a length in this way, in which case one specifies the conversion by giving the standard speed and a standard unit of length.

In fact, we do exactly that any time we draw a trend line on a piece of paper. The horizontal axis represents time, and each horizontal distance a certain amount of elapsed time. We shall even do so below on occasion. For the purposes of theoretical physics, however, we wish to invoke the least-time principle that simplifies so much of classical physics: A physical process evolves in such a way that the integral of the difference between kinetic and potential energy *with respect to time* is "stationary" (usually a minimum). That is why we shall generally homogenize dimensions and express them all as time. From Chapter 4 on, this aspect of the theory moves to center stage, and we shall then exclusively write intervals as time intervals.

Actually, we can measure the "distance" from one place to another in many different ways. For an astronaut, the distance between two points of space might be most practically measured as the amount of fuel required to get from one to the other. For an economy-minded ordinary citizen, the distance from, say Boston to Chicago, might be measured by the cost of the airline fare for a round-trip journey (in which case, many mutually inconsistent measures of the distance would exist).

The important aspect of Newtonian space-time to be kept in mind is that it involved *two* conventional standard units. There is no "natural" unit of time, and there is no "natural" unit of speed. The choice of each is arbitrary. We can think of the standard v_0 as the maximum rate at which information can be transmitted, calling it the "speed of information." This fact is in complete accord with the well-known fact that there is no natural unit of length in a Euclidean space. It is this "flatness" of Euclidean space that makes it possible to build scale models of vintage automobiles, ships, airplanes, and shopping malls, in which all lengths are shrunk in the same proportion and all angles are the same in the model as in the original. This is not possible, for example, when one tries to draw a map of a large portion of the Earth's surface. Changing the lengths also causes the angles to change. In fact, a sphere *does* have a natural unit of length, namely its radius. Similarly, the curved plane of hyperbolic geometry has a natural unit of length, which might be (for example) the distance at which the angle of parallelism is half of a right angle. (See Appendix 1 for details. Angles have an absolute meaning in all geometries, a right angle, for example, being exactly one-fourth of a complete rotation.)

In the special theory of relativity, by way of contrast, the constancy of the speed of light provides a natural link between space and time that is absent from the Newtonian model. It has profound—and observable—consequences for physics.

2.6. Absolute spaces and parameter spaces. Finally, we introduce one more foreshadowing of the world of general relativity, in the form of completely general coordinate systems. Even in pure mathematics a point $\boldsymbol{x}$ in n-dimensional Euclidean space $\mathbb{R}^n$ can often be usefully represented by *some other* n-tuple of real numbers different from its components. For example, in linear algebra it is often useful to change the basis of a vector space so that linear operator will have a simpler matrix (diagonal if possible). When we make a change of basis, the *coordinates* of a point in the absolute space are no longer equal to its *components*. The point *actually is* its components, but we may prefer to work with its coordinates. Thus, the space of components is the absolute Euclidean space $\mathbb{R}^n$ of n dimensions, and the space of coordinates in a given basis is a parameter space of n dimensions. The components coincide with the coordinates in the natural basis of the space, which consists of the vectors $(1, 0, 0 \ldots, 0, 0)$, $(0, 1, 0, \ldots, 0, 0)$,..., $(0, 0, 0, \ldots, 1, 0)$, $(0, 0, 0, \ldots, 0, 1)$.

Given that observers use parameter spaces, the problem of communication between them becomes important. How do they know when they are talking about the same physical quantity? What does it even mean for it to be "the same quantity" if it has no absolute definition? If the quantity is the star Sirius or the Atlantic Ocean, no real problem arises. But what if the two observers are talking about the intensity of a magnetic field? How do they know which measurements by one observer correspond to given measurements by the other? Indeed, using Newtonian mechanics, two observers in uniform relative motion do agree about magnetic fields, but not about electric fields. The "necessary evil" of parameter spaces, which Carl Friedrich Gauss (1777–1855) made into a convenience, brings with it the problem of reconciling parameters between observers. This is the problem of invariance, which has to be taken into account in general relativity (Chapters 5 and 6) and will be discussed in detail in Appendix 6.

Each observer uses an individual set of parameters to compute various important geometric objects, such as length, area, and curvature. Once we agree on a way for two observers to identify the parameter values corresponding to points, functions, and vectors, we are still left with the problem of reconciling the *processes* by which these objects are computed. If Observer O combines objects a, b, and c according to some algorithm to produce an object d, then Observer O' can interpret these objects as a', b', c', and d'. The question that naturally arises is: Suppose O' combines a', b', and c' following what is verbally the same algorithm that O followed? Will the result be the corresponding d'? The general answer is affirmative, provided all the objects that are combined are tensors. Detailed discussion of the problem is given in Appendix 6. Just to make this matter seem a bit less abstract, we note that the objects we are interested in are all obtained from the coordinate parameters through algebraic operations and differentiation. The criterion for an object to be a tensor is, in informal language, that when space-time coordinates are changed, its coordinates transform according to the chain rule for differentiation.

3. An Asymmetry in Newtonian Mechanics: Electromagnetic Forces

> *It is recognized that Maxwell's electrodynamics, as currently interpreted, leads to asymmetries that do not appear to be intrinsic to the phenomena when applied to moving bodies. Consider, for example, the electrodynamic interaction between a magnet and a conductor. Here the observable phenomenon depends only on the relative motion of the conductor and magnet, while in the usual interpretation a distinction is made as to which of them is moving. In particular, if the magnet moves and the conductor remains at rest, an electric field having a definite quantity of energy arises around the magnet, producing a current in regions where the conductor is present. But if the magnet is at rest and the conductor moves, no electric field arises around the magnet; on the other hand, an electromotive force arises in the conductor, not corresponding to any energy, but which nevertheless—assuming the equality of the relative motion in the two cases—produces electric currents of the same magnitude as those generated by the electric field in the first case.*
>
> Einstein ([**17**], p. 891).

When vector analysis was invented in the late nineteenth century, it seemed that a very powerful mathematical language had been created, one ideally suited to the purpose of getting compact expressions of physical laws. Now vector analysis is indeed a powerful tool, but in the first decade of the twentieth century it was a recent creation and by no means universally used. Einstein did not use the vector operations of curl and divergence in his 1905 paper on special relativity, although he did use the word *vector*. But he wrote out the Maxwell equations connecting electric and magnetic fields componentwise. His use of this seemingly more cumbersome notation may have been caused by the fact that two observers in relative motion do not agree that they are both using Euclidean geometry, in which the curl and divergence have a coordinate-free meaning.

Because Newton's second law of motion asserts that forces are directly proportional to acceleration, it follows that two observers in relative motion at constant velocity (zero relative acceleration) must agree about the magnitude and direction of all forces. This principle does not appear to have raised any doubts until the discovery of Maxwell's four laws of electromagnetism, when an asymmetry arose for two such observers looking at a charged particle moving in a pair of electric and magnetic fields. The two observers, it turned out, would agree about the magnetic field, but not about the electric field. They agreed about the *magnitude and direction* of the forces on the particle, but not about the *physical nature* of those forces. It was this asymmetry that Einstein remarked upon in his fundamental 1905 paper on special relativity. He made only a casual allusion to the famous Michelson–Morley experiment[12] of 1887 that had failed to detect any dependence of the speed of light on its direction of motion in a hypothetical absolute space. Einstein did, later on, discuss an 1851 experiment by Armand Hippolyte Louis Fizeau (1819–1896) of which the Michelson–Morley experiment was an improved reconstruction.

In the Newtonian system, there is a time axis common to all observers and a three-dimensional Euclidean space also common to all observers. When two observers wish to communicate their observations to each other, it is only necessary for one of them to say what coordinates are assigned to three points in space relative to that observer's origin and what event marks the epoch (time 0) of the time axis at that origin. The physical anomalies mentioned above, however, forced a reformulation of mechanics, in which time and space could not be separated. What Observer O takes to be an orthonormal coordinate system in space is not orthonormal as seen by Observer O'. As a result of that stark difference, each individual observer might at first sight seem to be isolated in a set of time and space coordinates, which are for that observer just like the old Newtonian ones, but does not agree with the equally valid system of time and space coordinates used by another observer. Two observers in relative motion almost appear to be inhabiting parallel universes. How can they determine whether an event occurring at point $\boldsymbol{x}$ at time t in the coordinates used by O is to be regarded as the event occurring at point $\boldsymbol{x}'$ at time t' in the coordinates used by O'? What common observations will enable them to make such an identification?

The key to solving this problem is their common line of motion, assuming that each believes the other is moving along a straight line at constant speed. We assume that each can at the very least observe the origin used by the other—it is convenient to think of the observers as "sitting" at their respective origins at all times—and assign a location to the origin of that other observer at any given time. We can identify the line in O-coordinates joining the two origins with the line in O'-coordinates joining the two origins.

After these lengthy preliminaries, we are at last ready to tackle the problem posed by the constancy of the speed of light and thereby explain the impossibility of getting 100% agreement on the order of events between two observers in relative motion.

4. The Lorentz Transformation

The starting point for the reformulation of mechanics is the assumption that the speed of light is a universal constant for all observers, independent of the motion of the observer or the source of the light. This assumption requires us to revise a number of intuitive notions about the order in which events occur.

An *event* is a point $(t; x, y, z)$ in $\mathbb{R}^4$, written with a semicolon to distinguish the time coordinate t from the three spatial coordinates (x, y, z).

[12]Named after Albert Abraham Michelson (1852–1931) and Edward Williams Morley (1838–1923). For details, see Chapter 8.

When we discuss events, we must keep in mind that their temporal order may depend on the observer. If the observers are not in relative motion, however, that is, the spatial coordinates of all points are constant in both frames of reference and the origins always coincide, then we can assume that the time coordinate of any event is also the same for both, and coordinates can be converted as in classical mechanics, that is, by a linear transformation $(t; x, y, z) \mapsto (t'; x', y', z')$ given by

$$\begin{aligned} t' &= t\,, \\ x' &= a_{11}x + a_{12}y + a_{13}z\,, \\ y' &= a_{21}x + a_{22}y + a_{23}z\,, \\ z' &= a_{31}x + a_{32}y + a_{33}z\,, \end{aligned}$$

where the matrix

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

is the invertible matrix that transforms coordinates in one fixed basis of $\mathbb{R}^3$ to another. In our case, it will always be a rotation matrix, since we are going to assume that our observers use only right-handed orthonormal bases in their coordinates.[13]

To see what these coordinate changes look like in relativity, imagine two observers, O and O', whose spatial frames of reference are moving relative to each other in a fixed direction at a constant speed u. Each of our observers is imagined to have a clock measuring time in the Newtonian way through some physical process that is, by definition, said to be proceeding at a uniform rate, and each has measuring instruments that measure distances and angles in such a way that triangles obey the trigonometric laws of Euclidean geometry. These are the *proper time* and *proper space* for that observer, and they have the properties of Newton's absolute time and space, including the property that time and space measurements are independent variables. The difference from Newton's system is that these times and spaces are not in agreement with the proper times and spaces of other observers. When two observers try to reconcile their measurements, each finds the space and time measurements of the other are *entangled* and so no longer appear to be independent variables when compared with his own. In particular, distances (x) along the common line of motion are mixed up with time (t), and only the difference $(ct)^2 - x^2$ (where c is the speed of light) is agreed upon by both observers.

For simplicity, we assume that at some instant of time, given the value $t = 0 = t'$ by both observers, the three mutually perpendicular coordinate axes used by O coincide with those used by O', and that the relative motion is a translation along the direction of the common x-axis (x'-axis) at constant speed. We assume that O''s origin is moving in the positive direction of the common axis at speed u (from O's point of view). Of course, from O''s point of view O's origin is moving along the x'-axis in the negative direction, that is, at speed $-u$. Because of the assumption

[13]As Einstein remarked in one of his popular expositions of relativity theory, it is impossible to define what a right-handed system is intrinsically. One can divide systems of orthonormal bases into two equivalence classes and always say whether two bases belong to the same class—the determinant of the transition matrix between coordinates in the two bases is positive—but otherwise, there is nothing intrinsic to either system that marks it as being "right-handed." Thus, what we are really saying is that we assume the coordinate transformation between any two observers has a positive determinant.

that the speed of light must be the same for both observers, we cannot now assume that they are using the same time coordinate, or that simultaneity means the same thing for both of them. The best we can assert is a kind of homogeneity in events, expressed by assertions like *if event P took place twice as far away from O's origin as event Q, and after an elapsed time (measured from the instant when the two origins coincided) twice as large as the time elapsed when Q occurred, as seen by O, then the same should be true from O''s point of view.* That is, if all four of the space-time coordinates of event P are twice those of event Q in O's system, the same should be true in O''s system.

Einstein must have had something like this in mind when he asserted that, because of our beliefs about the nature of time and space, it seems clear that the coordinates of an event in one frame of reference must be linear functions of those in the other frame. In other words, we are assuming that there is a linear transformation such that

$$\begin{aligned} t' &= a_{11}t + a_{12}x + a_{13}y + a_{14}z, \\ x' &= a_{21}t + a_{22}x + a_{23}y + a_{24}z, \\ y' &= a_{31}t + a_{32}x + a_{33}y + a_{34}z, \\ z' &= a_{41}t + a_{42}x + a_{43}y + a_{44}z\,. \end{aligned}$$

Our first assumption is that the yz-plane coincides point by point with the $y'z'$-plane at a time we shall take as the epoch (time 0) for both observers. That is, if $t = 0 = t'$ and $x = 0 = x'$, then $y' = y$ and $z' = z$. This assumption yields the equalities

$$\begin{aligned} 0 &= a_{13}y + a_{14}z\,, \\ 0 &= a_{23}y + a_{24}z\,, \\ y &= a_{33}y + a_{34}z\,, \\ z &= a_{43}y + a_{44}z\,. \end{aligned}$$

If these equalities are to hold for all y and z, then we must have $a_{13} = 0 = a_{14}$, $a_{23} = 0 = a_{24}$, $a_{33} = 1$, $a_{34} = 0$, $a_{43} = 0$, $a_{44} = 1$. The equations now read

$$\begin{aligned} t' &= a_{11}t + a_{12}x\,, \\ x' &= a_{21}t + a_{22}x\,, \\ y' &= a_{31}t + a_{32}x + y\,, \\ z' &= a_{41}t + a_{42}x + z\,. \end{aligned}$$

The assumption that the motion is along the x-axis in both systems implies that this axis is the same for both at all times. In other words, if $y = 0 = z$, then $y' = 0 = z'$ also. Putting these values in the last two equations, we find

$$\begin{aligned} 0 &= a_{31}t + a_{32}x\,, \\ 0 &= a_{41}t + a_{42}x\,. \end{aligned}$$

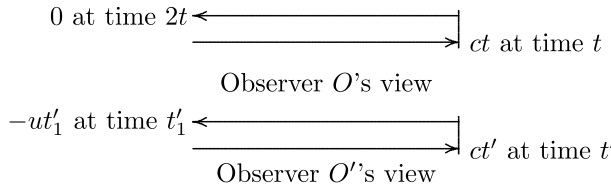

FIGURE 1.1. Top: A light ray traveling from the origin of Observer O's coordinate system to a mirror on the axis of common motion and then back to the origin, as described by O. Bottom: The same process as described by O'.

Since these equations hold for all t and x, we must have $a_{31} = 0 = a_{32}$ and $a_{41} = 0 = a_{42}$. Our transformation equations now read

$$\begin{aligned} t' &= a_{11}t + a_{12}x\,, \\ x' &= a_{21}t + a_{22}x\,, \\ y' &= y\,, \\ z' &= z\,. \end{aligned}$$

These determinations are not particular to the theory of relativity; they are independent of any assumptions about the speed of light c. In order to determine the four remaining coefficients a_{ij}, $i, j = 1, 2$, we need to introduce the assumption that c is the same for all observers. With that assumption, we first consider coordinates assigned to events on the axis of relative motion. Consider a light ray that leaves the common origin at time $t = 0 = t'$, travels to a mirror on the positive x-axis (which is also the positive x'-axis), arriving at time t according to Observer O, then is reflected straight back to the origin of O's system, necessarily arriving there at time $t_1 = 2t$ according to O. To Observer O', the light ray arrives at the mirror at some time t', then returns to O's origin at some later time t'_1, when that origin has coordinates $(-ut'_1, 0, 0)$. Thus we have two events, the arrival of the light ray at the mirror, to which the two observers assign coordinates $(t, ct, 0, 0)$ and $(t', ct', 0, 0)$ respectively, and its return to O's origin, to which the two observers assign coordinates $(2t, 0, 0, 0)$ and $(t'_1, -ut'_1, 0, 0)$, so that

$$\begin{aligned} t' &= a_{11}t + a_{12}ct\,, \\ ct' &= a_{21}t + a_{22}ct\,, \\ t'_1 &= 2a_{11}t\,, \\ -ut'_1 &= 2a_{21}\,t. \end{aligned}$$

The last two equations here imply that $a_{21} = -ua_{11}$.

Now let us look more closely at the return portion of this trip in the "light" of the fact that the speed of light c is the same for both observers. According to O', after leaving the mirror, the light traveled a distance $c(t'_1 - t') = ct' + ut'_1$. (Here the left-hand side represents the distance traveled as speed times time elapsed. The right-hand side represents it directly as the difference of the two distances from O''s origin to the starting point ct' and ending point $-ut'_1$.) Solving this relation for t', we find

$$t' = \frac{c-u}{2c} t'_1 \,.$$

Substituting this value of t' into the first two equations of the coordinate transformation, we find

$$\begin{aligned} \frac{c-u}{2c} t'_1 &= a_{11} t + c a_{12} t \,, \\ \frac{c-u}{2} t'_1 &= -u a_{11} t + c a_{22} t \,. \end{aligned}$$

Since $\frac{1}{2} t'_1 = a_{11} t$, we can cancel this factor; and we get, upon dividing the second equation by c,

$$\begin{aligned} \frac{c-u}{c} &= 1 + \frac{c a_{12}}{a_{11}} \,, \\ \frac{c-u}{c} &= -\frac{u}{c} + \frac{a_{22}}{a_{11}} \,. \end{aligned}$$

We now rewrite these equations as

$$\begin{aligned} \frac{a_{12}}{a_{11}} &= \frac{c-u}{c^2} - \frac{1}{c} = -\frac{u}{c^2} \,, \\ \frac{a_{22}}{a_{11}} &= 1 \,. \end{aligned}$$

Now, letting $\alpha = a_{11}$, we have $a_{21} = -u\alpha$, $a_{12} = -\frac{u}{c^2}\alpha$, and $a_{22} = \alpha$. Thus, we have the transformation

$$\begin{aligned} t' &= \alpha\Big(t - \frac{u}{c^2} x\Big) \,, \\ x' &= \alpha(-ut + x) \,. \end{aligned}$$

It remains to determine the factor α. This time we imagine a point fixed on O's y-axis. A light ray again leaves the common origin at time 0 (in both coordinate systems) and travels to this point, arriving at time t. Then O assigns to this arrival the coordinates $(t, 0, ct, 0)$, and, by what we know of the transformation so far, O' assigns coordinates, $(\alpha t, -\alpha u t, ct, 0)$ to this same event. For O', the event is the arrival at the point $(-\alpha u t, ct, 0)$ of a light ray that left O''s origin at time 0, and this arrival occurs at time αt. Since the proper space of each observer is Euclidean, we deduce that

$$c\alpha t = \sqrt{(\alpha u t)^2 + (ct)^2} = t\sqrt{(\alpha u)^2 + c^2} \,.$$

Canceling t, squaring the equation, and then solving for α, we find

$$\alpha = \frac{c}{\sqrt{c^2 - u^2}} = \frac{1}{\sqrt{1 - \frac{u^2}{c^2}}} \,.$$

The complete set of transformation equations is now best stated as a theorem.

THEOREM 1.1. *For two systems of measuring time t and t' and rectangular space coordinates (x, y, z) and (x', y', z') for which* (1) *the x-axis and the x'-axis coincide at all times and* (2)*, the origin of the primed system is moving along the* *along the x-axis at speed u, the four coordinates in the two systems are related by the following system of equations:*

$$t' = \alpha\Big(t - \frac{u}{c^2}x\Big)\,, \tag{1.1}$$

$$x' = \alpha(-ut + x)\,, \tag{1.2}$$

$$y' = y\,, \tag{1.3}$$

$$z' = z\,. \tag{1.4}$$

We shall refer to this transformation as the *Lorentz transformation* and say that it corresponds to a velocity vector $\boldsymbol{u} = u\boldsymbol{i}$ of O' relative to O. The reader can easily compute that the *space-time interval* between two events is the same for both observers:

$$\Delta s^2 = \Delta t'^2 - \frac{1}{c^2}\Delta x'^2 - \frac{1}{c^2}\Delta y'^2 - \frac{1}{c^2}\Delta z'^2 = \Delta t^2 - \frac{1}{c^2}\Delta x^2 - \frac{1}{c^2}\Delta y^2 - \frac{1}{c^2}\Delta z^2\,.$$

The interpretation of Eqs. 1.1–1.4, is that the event recorded by observer O as $(t; x, y, z)$ is the same event that is recorded by O' as $(t'; x', y', z')$, provided the two sets of coordinates are related by these equations (and the x-axis/x'-axis is the line of mutual motion). These equations make it possible for two observers to agree on what happens, say, to a particle that both observe moving along trajectories $\big(x(t), y(t), z(t)\big)$ and $\big(x'(t'), y'(t'), z'(t')\big)$ (where of course, the primes do *not* mean differentiation).

REMARK 1.3. It is sometimes useful to have a four-dimensional space-time in which all the coordinates have the physical dimension of length, (see Problem 1.19 below). For that reason, we shall occasionally replace the time coordinates t and t' by the "spatialized" times $\tau = ct$ and $\tau' = ct'$. When that is done, the matrix of the Lorentz transformation takes on a more symmetric appearance:

$$\begin{aligned}
\tau' &= \alpha\Big(\tau - \frac{ux}{c}\Big)\,,\\
x' &= \alpha\Big(-\frac{u\tau}{c} + x\Big)\,,\\
y' &= y\,,\\
z' &= z\,.
\end{aligned}$$

Thus, the Lorentz transformation imposes two significant modifications on Newtonian space-time:

(1) It provides a natural absolute velocity—the speed of light, denoted c—by which we can "temporize" spatial coordinates through Minkowski's "mystical formula." (See the note on p. 5.) The arbitrary velocity v_0 we used earlier for this purpose is no longer arbitrary, being replaced by c. The unit of time remains arbitrary, however.

(2) It does not preserve the Euclidean metric on $\mathbb{R}^4$, in which a point $(t; x, y, z)$ has the square-norm $t^2 + x^2 + y^2 + z^2$; instead, it preserves the pseudo-metric given by the bilinear form $t^2 - (x^2 + y^2 + z^2)/c^2$, which assumes both positive and negative values. This bilinear form, used as a metric, creates a *pseudo-Euclidean* space. This space is still flat, since its metric coefficients (the coefficients in the bilinear form) are still constant, but the square of the space-time interval between two events can be negative. Such intervals are called *spacelike* because the spatial portion of the proper

time interval is larger than the time portion; intervals in which the time portion is larger are called *timelike.*

Those who appreciate the power and beauty of vectors will want to state it in vector form, and this can be done, by simply decomposing the spatial portion of the vectors in $\mathbb{R}^4$ into vectors parallel to $\boldsymbol{u}$ and perpendicular to $\boldsymbol{u}$, that is, writing

$$\boldsymbol{x}' = \frac{\boldsymbol{x}' \cdot \boldsymbol{u}}{\boldsymbol{u} \cdot \boldsymbol{u}}\boldsymbol{u} + \Big(\boldsymbol{x}' - \frac{\boldsymbol{x}' \cdot \boldsymbol{u}}{\boldsymbol{u} \cdot \boldsymbol{u}}\boldsymbol{u}\Big)\,.$$

Since the time coordinate is easily written in terms of the vector dot product, the resulting vector equation is:

$$(t'; \boldsymbol{x}') = \left(\alpha\Big(t - \frac{\boldsymbol{u} \cdot \boldsymbol{x}}{c^2}\Big); \boldsymbol{x} + \Big((\alpha - 1)\frac{\boldsymbol{x} \cdot \boldsymbol{u}}{\boldsymbol{u} \cdot \boldsymbol{u}} - \alpha t\Big)\boldsymbol{u}\right),$$

where $\alpha = c/\sqrt{c^2 - \boldsymbol{u} \cdot \boldsymbol{u}}$. The inverse relation is

$$(t, \boldsymbol{x}) = \left(\alpha\Big(t' + \frac{\boldsymbol{u} \cdot \boldsymbol{x}'}{c^2}\Big); \boldsymbol{x}' + \Big((\alpha - 1)\frac{\boldsymbol{x}' \cdot \boldsymbol{u}}{\boldsymbol{u} \cdot \boldsymbol{u}} + \alpha t'\Big)\boldsymbol{u}\right),$$

Here, for the first but not the last time, we emphasize that both observers *must* be using $\boldsymbol{i} = \boldsymbol{u}/u$ as the first vector of an orthonormal basis in order for the vector equation to be valid. These must be such that the coordinates of the velocity vector $\boldsymbol{u}$ that O ascribes to O', are the negatives of the coordinates of the velocity that O' ascribes to O. Only in that sense is it helpful to say that the velocity of O relative to O' is $-\boldsymbol{u}$. The Lorentz transformation was derived assuming the two observers are using such coordinates. If either of them rotates the spatial axes, this vector equation becomes false when written out in terms of components in the two systems. When we introduce a third observer O'' having a velocity $\boldsymbol{v}$ relative to O'—we assume its components are assigned by O'—that is not parallel to $\boldsymbol{u}$, we shall find that the relative velocity vectors $\pm\boldsymbol{w}$ between O and O'' do not have this property unless all three of the observers rotate their spatial coordinates. Composing two Lorentz transformations is thus not just a simple matter of multiplying the matrices of the two transformations in fixed coordinate systems used by the three observers. If the three observers are using arbitrary orthonormal coordinate systems, the complete composition requires multiplying five matrices rather than two.

Thus, although the vector notation is compact, it must be used carefully. We remark that the quantities $\boldsymbol{u} \cdot \boldsymbol{u} = u^2$ and $\alpha = c/\sqrt{c^2 - u^2}$ determine each other ($\boldsymbol{u} \cdot \boldsymbol{u} = u^2 = c^2(\alpha^2 - 1)/\alpha^2$), so that we could eliminate one of the two from the Lorentz transformation, replacing $(\alpha - 1)/(\boldsymbol{u} \cdot \boldsymbol{u})$ by $\alpha^2 c^2/(\alpha + 1)$.

5. Contraction of Length and Time

Now that we have introduced the Lorentz transformation and the pseudo-Euclidean space-time that it operates on, we would like to analyze the effect of this transformation on time and space separately. For this purpose, we can suppress the y and z-coordinates of our two observers, since they are the same. We imagine two situations, in one of which O' observes the clock used by O over a period of time and in the other of which O' observes the distances that O assigns to points along their common line of motion. In each case, O' wants to compare O's measurements with his own. To that end, it is convenient to imagine that each observer has set up a post at every point along the common line of motion with a

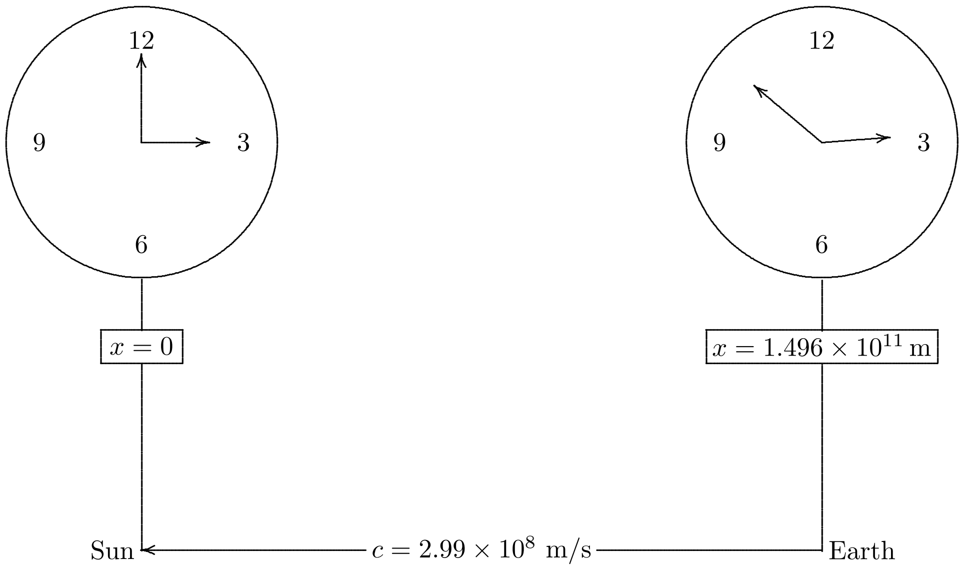

FIGURE 1.2. An observer located at the Sun uses a telescope to observe a clock on the Earth synchronized with his own. Left: The clock on the Sun. Right: The image of the clock on Earth observed at the same instant on the Sun. This image requires about 8.5 minutes to reach the Sun.

sign on the post giving its distance from the origin and a clock showing the time. Since each observer is using the familiar Newtonian ideas of space and time, it is possible for each to synchronize all the clocks in the system. The way to do so is to imagine, say O seated at his own origin holding a clock, and looking through a telescope at one of these sign posts, say corresponding to coordinate x. If O's clock reads time t, then, looking through the telescope at the signpost located at x, O should see the clock showing time $t - |x|/c$, to account for the fact that the light revealing the clock to O left the sign post earlier and required time $|x|/c$ to reach the origin. If the two clocks show times related in this way, then the clock at x can be regarded as synchronized with the clock at the origin. In that way, O gets a concept of "now" that applies all over the universe, although what is happening "now" at a distant location cannot be known "now." That information must await the arrival of a signal from the point with coordinate x, and cannot occur until time $|x|/c$ after "now." Observer O' can perform a similar synchronization, and we shall always assume that both observers have already equipped every point of space with such a sign post bearing an accurate clock. The synchronization procedure is illustrated in Fig. 1.2.

The problem is that, while the x-axis is the same for both observers, those sign posts are moving past each other. Although posts with negative x-values are in locations coinciding with posts bearing negative x'-values, a post with a negative x' value will eventually reach O's origin and move to a position coinciding with a post bearing a positive x-value. As it will turn out, the two observers will assign different x-coordinates and different clock times to all points on the x-axis except the origin itself, even at the time both call $t = 0 = t'$. We need to see what the discrepancy amounts to. We shall see that, at the very least, it amounts to the impossibility of agreement on what is happening "now" at time $t = 0 = t'$ anywhere except at the common origin of the two systems, which coincide at that time—or should we say, those times, since time is something they don't generally agree on.

5.1. Time contraction. Imagine O' reading the time t on the clock located at O's origin and comparing its reading with his own clock. Since O's origin at what O' calls time t' is located at $(-ut', 0, 0)$—that is, the event $(t; 0, 0, 0)$ to O is the event $(t'; x', 0, 0) = (\alpha t, -\alpha ut, 0, 0)$ to O'—we have the equality $t' = \alpha t$. Since $\alpha > 1$, O' perceives O's clock as "running slow." That is, O', reading the clock at O's origin and comparing it with his own, says that it assigns too small a value to the interval between events occurring at O's origin. There is perfect reciprocity here: O observes a clock located at O''s origin as running slow by the same factor $1/\alpha = \sqrt{1 - u^2/c^2}$. If these two facts appear to contradict each other, note that they refer to different sequences of "events." In the first case, we are looking at the event that O describes as $(t; 0, 0, 0)$ and computing the time coordinate assigned to it by O', which is $t' = \alpha t$. In the second case, we are looking at the event that O' describes as $(t'; 0, 0, 0)$. Thus we are describing events at a location regarded as fixed by one of the observers but not the other in each case.

It should not be thought that the apparent slowness each observes in the other's clock is caused by the fact that each observer is moving away from the other. The same time contraction is observed even when they move toward each other. That is, it holds for negative times t as well as positive ones. True, given that the two clocks agree when the origins coincide, O' will observe a later time on O's clock than on his own as the two observers approach each other and an earlier time as they recede from each other. The time elapsed between two *events* occurring at O's origin—say, for example, the time between two flashes of a beacon located at that origin—will be shorter as read by O' from O's clock than as read from O''s own clock, whether the two are getting closer to each other or farther away.

Nor is the difference due to the fact that the light from O's origin takes some time to arrive at O''s origin. To correct for that time lag, we apply the Newtonian synchronization discussed above. The relativistic computation assumes that O' has *already corrected for that time lapse* by keeping a clock synchronized with his own at every point of space and recording the time of an event at each point using that clock. For an event occurring at any given point P that is fixed in O''s frame, the time O' will assign to that event is the time t at which a light ray originating at P at the time of the event reaches O''s origin, less d/c, where d is the distance from the point P to O''s origin. Let us compare the time shown on O's clock when O passes through the point P with the time O' assigns to that event. In other words, we want to compare O's clock at the instant when O passes through the point P with the clock at P synchronized with the one at O''s origin. The two clocks will not show the same time when O passes through P. To see why, suppose O passes through P at the time t_0 in O's frame. Suppressing the unnecessary y and z coordinates, O assigns to this event the coordinates $(t_0, 0)$. To O', that event has coordinates $(t'_0, x'_0) = (\alpha t_0, -\alpha u t_0)$, and so O' assigns time αt_0 to that event. Since $\alpha > 1$, we conclude that *O records a smaller time interval between events occurring at his own origin than O' records.*

Now let us consider the arrival time of information from that event at O''s origin. That time will be

$$\alpha t_0 + \alpha|ut_0/c| = \begin{cases} t_0\sqrt{\frac{c+u}{c-u}} & \text{if } t_0 > 0\,, \\ t_0\sqrt{\frac{c-u}{c+u}} & \text{if } t_0 < 0\,. \end{cases}$$

We emphasize that Observer O' assigns this time to the *arrival of information* about the original event, but assigns to the original event at O's origin the time $\alpha t_0 = ct_0/\sqrt{c^2-u^2}$. The "Newtonian correction" for the speed of information is the difference in those two times, which is $u|t_0|/\sqrt{c^2-u^2} = \alpha u|t_0|/c$, precisely the time required for the light to travel from the location of the first event $(-\alpha ut_0)$ to the location of the second event (0).

5.2. The relativistic Doppler shift. To pursue this line of thought one step farther, we can imagine O broadcasting a signal in the form of a sine wave whose peaks are at times nt_0, $n = 0, \pm 1, \pm 2, \ldots$ and hence have frequency $\nu = 1/|t_0|$. Assuming $t_0 > 0$, we see that those peaks will reach O' at times $nt_0\sqrt{(c+u)/(c-u)}$ and hence the received signals will have frequency $\nu\sqrt{(c-u)/(c+u)}$. This is a "red shift" since the frequency of the received signals is smaller than the frequency of the transmitted signal. That of course is because the two observers are moving apart for positive values of time. The reader can verify that for negative values of t_0, when the two observers were approaching each other, the shift is a "blue shift" by the factor $\sqrt{(c+u)/(c-u)}$.[14]

In summary, we have two time intervals being measured by two different observers. There is the time between the transmission of the two peaks at O's origin, which O measures as $|t_0|$ and O' as $\alpha|t_0|$. And there is the time interval between the reception of the two peaks at O''s origin, which, if $t_0 > 0$, O' measures as $t_0\sqrt{(c+u)/(c-u)}$ and O as $\alpha t_0\sqrt{(c+u)/(c-u)} = ct_0/(c-u)$. (These two events both occur at the same point of space in O''s frame, and hence the time interval that O measures between them is α times the interval measured by O'; that follows from the symmetry of the principle enunciated above.) From O's point of view, the reception times differ by more than the transmission times because O' is moving away, and each successive peak has farther to travel in order to reach O'. In fact, O perceives the Doppler shift in the frequency of the signal O' is receiving to be by a factor $1 - u/c$ when the two observers are moving apart and by a factor $1 + u/c$ when they are moving toward each other. This is precisely the classical Doppler shift. To compare these quantities, we have the following inequalities for a signal

[14]This factor characterizes the *relativistic Doppler shift*, which differs from the acoustic Doppler shift that applies for signals transmitted as waves in a stationary medium. In the latter case, the shift in frequency is by the factor $1 \pm u/c$ for observer speeds u smaller than the speed c of the signal. The Doppler shift is named after the Austrian physicist Christian Doppler (1803–1853), who identified it in 1842.

broadcast by O and received by O' when the two are moving apart:

$$\begin{aligned}
\nu &= \frac{1}{t_0} = \text{measurement by } O \text{ of the broadcast frequency}\\
&> \nu\sqrt{1-\frac{u^2}{c^2}} = \text{measurement by } O' \text{ of the broadcast frequency}\\
&> \nu\sqrt{\frac{c-u}{c+u}} = \text{measurement by } O' \text{ of the received frequency}\\
&> \nu\Bigl(1-\frac{u}{c}\Bigr) = \text{measurement by } O \text{ of the received frequency}\,.
\end{aligned}$$

The analysis just given may prove helpful in working Problem 1.2 below.

5.3. Spatial contraction. If we imagine O' reading the coordinates that O assigns to the points on the x-axis at a given instant of time, say $t' = 0$ as judged by O', the second equation of the inverse Lorentz transformation says $x = \alpha x'$, and since $\alpha > 1$, it follows that $|x'| < |x|$. Thus O' perceives O's description of the distance between two points on the line of common motion as too large. Effectively, O' thinks O's "yardstick" has shrunk and is giving readings that are too large by the same factor by which he perceives O's clock to be assigning time intervals that are too small between events at O's origin. Notice that the conflict is over the time intervals between events transpiring at a fixed location in O's system and spatial intervals between points that are fixed in O's system. To phrase this fact another way, what O is calling one meter of length (x) is, according to O', actually less than a meter (x'): In particular, O', undertaking a journey between two fixed points in O's system, will consider the journey to be shorter than O considers it.

A symmetric phenomenon arises when O measures the time intervals between events at a location that is fixed in O''s coordinates or the spatial intervals between points that are fixed in O''s coordinates. Confusion arises unless one keeps firmly in mind whose coordinates are fixed. This caution is needed especially in connection with the twin paradox, discussed below.

Length contraction along the line of mutual motion has an interesting effect on the measurement of volume:

THEOREM 1.2. *The volumes V_0 and V of a solid body B measured respectively in a system of coordinates at rest with respect to B and one in motion at constant speed u with respect to B are related by*

$$V = V_0\sqrt{1-\frac{u^2}{c^2}}\,.$$

PROOF. For any cube with sides parallel to the axes, one of the sides shrinks by the factor $\sqrt{1-u^2/c^2}$ when measured by the moving observer. The other two stay the same, so that the cube as measured by the moving observer is a parallelepiped and the volumes of the cube and parallelepiped are related as stated in the theorem. Since any volume can be approximated with arbitrary precision by a union of such cubes, this relation must hold for all volumes. □

6. Composition of Parallel Velocities

Suppose O' describes the passage of a particle from a point a on the x'-axis at time t'_0 to point b on the same axis at time $t'_1 > t'_0$. Then O' will say the journey required time $t_1' - t_0'$ and covered the distance $b - a$. In contrast, O will say that the particle passed from the point $\alpha(a + ut_0')$ at time $t_0 = \alpha(t_0' + ua/c^2)$ to the point $\alpha(b + ut_1')$ at time $t_1 = \alpha(t_1' + ub/c^2)$, that the passage required time $t_1 - t_0 = \alpha(t_1' - t_0' + (b-a)u/c^2)$ and covered distance $\alpha\bigl(b - a + u(t_1' - t_0')\bigr)$. In terms of the speed the two observers ascribe to the particle, O' thinks it is $v = (b-a)/(t_1' - t_0')$, while O thinks it is $w = \bigl(u(t_1' - t_0') + (b-a)\bigr)/\bigl((t_1' - t_0') + u(b-a)/c^2\bigr)$. Replacing $b-a$ by $v(t_1' - t_0')$ and cancelling $t_1' - t_0'$, the reader can easily verify the fundamental equation

$$w = (u + v)/(1 + uv/c^2)\,.$$

This formula gives the velocity of the particle relative to O, given that its velocity relative to O' is v and O''s velocity relative to O is u. It works only in the particular case when the two velocities are parallel.

Remark 1.4. The alert reader will have observed the similarity of the formula

$$w = \frac{u + v}{1 + \frac{uv}{c^2}}$$

to the addition formula for the hyperbolic tangent function:

$$\tanh(x + y) = \frac{\tanh(x) + \tanh(y)}{1 + \tanh(x)\tanh(y)}\,.$$

This similarity suggests a useful mapping from the space of relativistic velocities along a line, which can be thought of as the interval $(-c, +c)$ (negative values meaning motion to the left and positive values to the right, as seen by a fixed observer), to the real line $(-\infty, +\infty)$, namely the mapping $u \mapsto \operatorname{arctanh}(u/c)$. If we introduce a unit of length k, the function $U = k \operatorname{arctanh}(u/c)$, whose inverse is $u = c\tanh(U/k)$, associates velocities with lengths, and the speed c would correspond to an infinite length. Before we develop this idea further, we wish to make a few comments about the use of numbers to represent measurable quantities that are continuous, such as distance, time, mass, velocity, acceleration, force, pressure, and many others.

Mathematical functions do not accept what we call concrete numbers or variables—those with a geometric or physical dimension attached to them—as input. When we develop a function as a power series, for example,

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots\,,$$

all the terms must have the same geometric/physical dimension. Otherwise, they cannot be added. If the coefficients a_n are all to be pure numbers, then x must also be a pure number. Thus, when we give the value of a continuous physical quantity, what we are really giving is its ratio to a *standard unit* of that category. In the present context, we are talking about speed, and in the MKS system, the standard unit of speed is one meter per second. In relativity, however, c, the speed of light serves as a much more natural standard of speed, since it is the same for all observers. When we substitute the speed u of an object into a formula, what we are really inserting, usually without thinking about it, is the ratio u/c, which is a dimensionless real number between -1 and 1.

With that dimensional consideration in mind, we are now going to make an association that will turn out to have interesting consequences later in this chapter. Since we like to draw pictures representing velocities as lengths, we shall associate with each speed u a length U, referred to a standard length k, making the following correspondence:

$$\begin{aligned} u &= c \tanh\left(\frac{U}{k}\right), \\ U &= k \operatorname{arctanh}\left(\frac{u}{c}\right) = k \ln \sqrt{\frac{c+u}{c-u}}\,. \end{aligned}$$

(See Problem 1.8 below.)

Under this correspondence we find that

$$\frac{u+v}{1+\frac{uv}{c^2}} \leftrightarrow c \tanh\left(\frac{U}{k} + \frac{V}{k}\right).$$

A further advantage of this convenient substitution is that it gives an elegant expression for the Lorentz magnification factor $\alpha = c/\sqrt{c^2 - u^2}$, namely $\alpha = \cosh(U/k)$. That relation could be used to simplify the study of relativistic velocities in two dimensions. For other reasons, however, we shall work laboriously through our transformations using only the language of velocities and only later (mostly in Appendix 1) reveal its astonishing connection with hyperbolic geometry.

Thus, under the pairing $u \leftrightarrow c \tanh(U/k)$ relativistic velocities (positive, negative, and zero) are in one-to-one correspondence with lengths, and the usual geometric addition of collinear lengths corresponds to the relativistic addition of the corresponding velocities. Moreover, the speeds $\pm c$ correspond to the points $\pm\infty$ on the extended real line.

7. The Twin Paradox

One of the interesting oddities of special relativity is popularly known as the *twin paradox.* To state it colorfully, we shall ask our twins Mary and John to reprise their roles in the 21$^{\text{st}}$ century, assuming the roles of our two observers O and O'. We shall let John represent O, the stay-at-home twin, and Mary the traveling astronaut O'. This being no longer the eighteenth century, they no longer communicate by letters sent across an ocean. Instead, they keep in touch by radio, making direct conversation practical for a little while. As the distance between them increases, however, messages take even longer to go from one to the other than they did in the eighteenth century, when Mary was on the surface of the Earth. As this paradox is stated, Mary leaves the Earth in a rocket ship bound for a planet orbiting a nearby star. Finding the planet unsuitable even for a vacation, much less as a place to settle down, she reverses direction immediately upon arrival and returns at the same uniform speed. Since the ship travels very fast, her brother John back home, observing the clock on the ship, sees it losing time in comparison with his own clock, both on the outward journey and on the return journey. (To visualize this effect, imagine that John has set a clock synchronized with his own at every point that Mary passes through. He can carry out this synchronization, since he and the clocks remain at rest relative to one another. If he watches Mary's journey through a telescope, he can directly compare the time on her clock with the time on the fixed clock as she passes it, and Mary's clock will indeed keep losing time.) Moreover, what is true of the clock on the ship is true of all physical and chemical processes, including the aging process in Mary. Thus, when she returns to the Earth, she really is younger than her twin brother.

If that were all there is to the paradox, we might be surprised, but not astonished. But, if we believe in the relativity of motion, we should be flabbergasted, or so it seems at first sight. After all, in the frame of reference used by Mary, it was her brother John who went traveling, along with the rest of the universe, and reversed directions just as the star arrived at her rocket ship. It was his clock that was running slow. Why isn't John the younger of the two?[15] This paradox has been discussed in many popular accounts of relativity. Quite often, the discussion invokes the general theory of relativity, according to which clocks do run slower in a gravitational field. However, this phenomenon was discussed by Einstein before he developed the general theory of relativity; and, as we shall show below, the fact that Mary winds up younger in this scenario follows from the Lorentz transformation alone and does not require the general theory or any invocation of inertial forces to explain it.

At the heart of the matter lies the FitzGerald–Lorentz contraction. The two siblings agree as to their relative speed, which we have called u and will assume constant here. But they do *not* agree on the distance between the point where they passed each other and the star that later came by Mary's ship. If John measures that distance as d, Mary measures it as d/α, where $\alpha = c/\sqrt{c^2 - u^2}$. This is the case on both the outward and return journeys, so that to Mary the journey covered a total distance $2d/\alpha$ and since she was traveling at speed u relative to the Earth-bound frame, the total elapsed time for the round-trip journey was $t = 2d/(\alpha u)$. That is the amount of proper time elapsed on the rocket ship during the round-trip journey, whether measured by a mechanical clock on board the ship or by the aging process in her body. But from John's point of view, Mary traveled a distance d and back at speed u, so that his clock records an elapsed time of $2d/u$. Since $\alpha > 1$, the twins will agree that Mary's clock shows less time elapsed than John's.

This scenario does not have the symmetry that is sometimes erroneously said to follow from the relativity of motion. To be sure, John also observes that the distance between two *fixed* points in Mary's frame of reference is shorter than Mary herself measures them to be. *But the star and Mary are not relatively fixed. The distance between them is constantly changing.* Taking the primary axis for both observers to be their line of mutual motion, so that the star lies on that axis, we see that when the siblings pass each other the first time, Mary is not setting out to reach the point on her own axis that John regards as being at distance d from Mary's origin at that instant. In fact, to make the story an accurate reflection of relativity, Mary can't reach *any point* other than the origin in her own space, since we imagine that is where she stays located. Her spatial coordinates move along with her. If the twins agree to synchronize their clocks at time 0 when they pass each other, the two clocks they are keeping at Mary's destination will not agree as to the time when Mary's journey began. Since the star is fixed as far as John is concerned, he will say it is also time 0 on the star, which is at distance d. In other words, if he has placed a clock at that star synchronized with his own and looks at it through a telescope, he will always see that clock showing time $t - d/c$, when his own Earth-bound clock shows time t.

[15]If you look on the Internet, you will not have any difficulty finding people, some even claiming to have advanced degrees in physics, who deny the paradox for exactly that reason. Such people ought to know better. The reality of the twin paradox has been confirmed by experiment.

Now, what John regards as the event $(t; x, y, z)$ is seen by Mary as the event $\big(\alpha(t - ux/c^2); \alpha(x - ut/c), y, z\big)$. John, as just remarked, has a clock at the distant star synchronized with his own. At the instant ($t = 0$) when Mary passes him, he records the event at the distant star as $(0, d, 0, 0)$, the event being that John's local clock at the star reads zero. Mary, however, will record this same event as $(-\alpha ud/c^2; \alpha d, 0, 0)$, so that in her coordinates John's clock at her destination read zero when the star was at the distance αd from her origin, and that was at the *earlier* time $-\alpha ud/c^2$. Between that earlier time and the time when she passed her brother, which both agree was time 0, the distance between Mary and the star shrank from αd to $\alpha d - \alpha u^2 d/c^2 = \alpha d(1 - u^2/c^2) = d/\alpha$; that is the distance she plans to travel.

True, Mary can regard herself as remaining in one place, at rest, while her brother moves at speed $-u$, but she does not measure the distance that John moved as d. When the twins meet for the second time, Mary will say that John traveled only the distance $2d/\alpha$, and her clock will show elapsed time $2d/(\alpha u)$, just as we found when analyzing the situation from John's point of view. The difference is that the star *did not move* in John's frame of reference, whereas it did move in Mary's frame of reference.

To summarize: Mary, whose distance to the turnaround point is changing over time, is the one who stays younger, because her biological clock is slow. As some writers express the matter, the traveling twin moves out of the frame of the Earth-bound twin, then moves back in again.[16] It is the bookkeeping involved in making those transitions that causes the loss of time for the traveling twin.

And yet... one still has the feeling of something-wrong-here. After all, according to Mary, the clock her Earth-bound brother uses has been running slow for the entire time of the journey. Why then does it show a *later* time when the twins meet for the second time? How can the tortoise, who always runs slower than the hare, nevertheless win the race? How can someone go into a revolving door behind you and emerge ahead of you? It seems to be a sleight-of-hand trick, like the conjuror who asks you to draw a card from a deck and hold onto it, then suddenly pulls it out of the deck that you drew it from. Somehow, when you weren't looking, the card got switched, but how?

To make the matter as plain as possible, "the card was switched" before the journey even started. In fact, Mary cannot synchronize her clock with *any* accurate clock on the star because the two are in relative motion. As long as we focus our attention on the Earth-centered frame of reference, we do not notice the switch, and then we are surprised when we look at Mary's clock and discover that it is running slow as she arrives at the star.

The algebraically simplified form in which we derived the Lorentz transformation requires a common event as origin, an event that can be regarded as having both space and time coordinates equal to zero in both systems. For this hypothetical journey, let O, using time-space coordinates (t, x), be John, and let O', using coordinates (t', x') be Mary. There are three events that we use as anchors here. Event 0, when Mary passes Earth heading to the star; Event 1, when she arrives at

[16]This statement assumes that Mary began and ended the journey standing on the Earth; it ignores the (perhaps tiny) portion of the journey during which acceleration was needed. We prefer to picture Mary as constantly whizzing around and just happening to pass her brother going in opposite directions at two different times, while hurtling past him. That way, we avoid any need to discuss acceleration.

the star and immediately heads back toward Earth at the same speed; and Event 2, when she passes Earth for the second time. We have to change origins at Event 1 in order to use Lorentz coordinate transformations on the return journey. For any event E occurring at John's origin at time t, we have O-coordinates $(t, 0)$ and O'-coordinates (t', x'), where $t' = \alpha t$ and $x' = -\alpha u t$. Thus, during the outbound portion of the journey Mary does record later times for all events at John's origin than John records for them. In that sense, Mary can say that John's clock is slow. But Event 1 *does not occur at John's origin.* We have focused our attention on the wrong place and by so doing missed the "card switch." To John, Event 1 has coordinates, say (t_1, d), where $d = ut_1$. According to the Lorentz transformation, Mary assigns coordinates (t_1', d') to that event, where $d' = \alpha(d - ut_1) = 0$—that is, it occurs at Mary's origin, as we already knew—and $t_1' = \alpha(t_1 - ud/c^2) = \alpha t_1(1 - u^2/c^2) = t_1/\alpha$, which is *smaller* than t_1. It is the ambiguity in the meaning of the term *simultaneous* that causes the surprise. By John's measurements, Event 1 occurred at time $t_1 = d/u$; by Mary's, it occurred at the earlier time $t_1' = t_1/\alpha$. It is the same event, of course, namely Mary's arrival at her destination.

Starting with Event 1, John and Mary both need to "zero out" their coordinate systems for Mary's return journey, and the velocity u needs to become $-u$. The Lorentz transformation that is in effect for this part of the journey is as follows:

$$\begin{aligned} t' - t_1' &= \alpha\Big((t - t_1) + \frac{u(x-d)}{c^2}\Big)\,, \\ x' &= \alpha\big(x - d + u(t - t_1)\big)\,. \end{aligned}$$

Then, when Event 2 occurs (Mary arrives back at Earth), John's coordinates for this event will be $t = 2d/u = 2t_1$, $x = 0$, so that Mary's will be $t' = t_1' + \alpha(t_1 - ud/c^2) = t_1/\alpha + \alpha t_1 - u^2 t_1 \alpha/c^2 = t_1\alpha(1/\alpha^2 + 1 - u^2/c^2) = 2t_1/\alpha$ and $x' = \alpha(-d + ut_1) = 0$. Thus, we can rigorously compute that the time on Mary's clock will be $1/\alpha$ times the time on John's clock: Mary will be younger by this factor.

Switching to our hare-and-tortoise analogy, we recall that the tortoise won the race because the hare took a nap. We are the ones who were caught napping in this race, focusing our attention on the Earth-bound clock, since the episode began at Event 0 and ended at Event 2, both of which took place on the Earth. We should have looked at the clocks at Mary's destination.

To phrase the matter in terms of the conjuring-trick analogy, the assertion that O's clock is running slow is the kind of distraction the conjuror uses to get you to focus on the card you thought was in your hand (John's clock on the Earth) but which was in fact *always* in the deck (John's clock on the star), not in your hand. The switching is revealed when your attention finally focuses instead on the deck of cards. If you would like to see the trick performed in slow-motion, work Problem 1.2 below.

8. Relativistic Triangles

With the advent of special relativity, the clean, simple Newtonian system, with its absolute, universal space and time, became untenable. It became impossible to determine the angle made by two lines at a given time, since observers in relative motion to each other may agree on the location of one point at a given time, while disagreeing as to the location of a second point at that same instant of time.

As a result, two observers will probably disagree as to what points constitute the vertices of a given triangle at a given time. Although each individual observer is using his own proper space, which is Euclidean, when that observer's findings are communicated to a second observer, the lines that the first observer regards as fixed are moving when viewed by the second observer, and that motion can cause perpendicular directions to go askew. The only kind of perpendicularity agreed upon by two observers moving at constant speed and in a constant direction relative to each other, is one particular system of coordinate axes consisting of the line of their mutual motion (that is, the line from each to the other), and the planes perpendicular to it. Even in that case, they do not agree about the unit of length along the first axis.

The absence of absolute simultaneity makes it difficult to discuss the "triangle whose vertices are P, Q, and R." While those points may be fixed in the spatial coordinates of Observer O, if O and O' are moving relative to one another, Observer O' will see a triangle of very different shape. Thus, triangles of position are slippery objects, impossible to define in an observer-independent way. What we are going to call *relativistic velocity triangles* are much better behaved in this regard.

We shall almost always find it easier to assume that suitable rotations of coordinates have already been performed by O and O', that is, that O and O' are already using the "privileged" coordinate systems in which they share the first spatial axis. A vector equation used by O to express a relationship between vectors measured only by O (for example, Maxwell's laws) is independent of the orthonormal basis used by O, since rotations preserve the vector operations of addition and scalar multiplication, as well as the dot and cross product on $\mathbb{R}^3$. But O *cannot* rotate axes and reinterpret correctly data received from O'. Transmitting the rotated data back to O' will not produce a rotation of the original data. We shall have occasion to write the laws of mechanics and electromagnetism in vector form and exhibit the transformation of those laws from one observer to another. Such equalities are not truly vector equalities, since they hold only when the vectors are expressed in the privileged bases of the two observers' systems, that is, those that share a common spatial axis along the direction of mutual motion.

As an example of what we have been talking about, notice that O regards the lines whose equations are $y = 2x$, $z = 0$ and $x = -2y$, $z = 0$ as mutually perpendicular. But at any given instant t', O' considers these to be the lines with equations $y' = 2\alpha x' + 2\alpha ut'$, $z' = 0$ and $\alpha x' + \alpha ut' = -2y'$, $z' = 0$, which are not perpendicular unless $u = 0$. They have slopes $m_1 = 2\alpha$ and $m_2 = -\alpha/2$, and the usual condition for perpendicularity ($m_1 m_2 = -1$) implies $\alpha = 1$, which is true only when $u = 0$.

8.1. Composition of nonparallel velocities. If O and O' are using the coordinate systems for which the Lorentz transformation equations were given, so that the velocity of O' relative to O is $\boldsymbol{u} = (u, 0, 0)$ and the velocity of O relative to O' is $-\boldsymbol{u} = (-u, 0, 0)$, and if an observer O'' whose origin also coincided with the origins of O and O' at time 0 is moving relative to O' with velocity $\boldsymbol{v}$ making an angle[17] η with $-\boldsymbol{u}$ according to O', so that O'''s origin at time t' is given by[18] O' as $(-vt'\cos\eta, vt'\sin\eta, 0)$, then the position of O'''s origin at time t' is interpreted by O as the event

$$\begin{aligned} t &= \alpha t'(1 - uv\cos\eta/c^2)\,, \\ x &= \alpha t'(u - v\cos\eta)\,, \\ y &= t'v\sin\eta\,, \\ z &= 0\,. \end{aligned}$$

We find the velocity $\boldsymbol{w}$ of O'' relative to O by dividing each of the spatial coordinates by t:

$$\begin{aligned} w_1 &= \frac{u - v\cos\eta}{1 - uv\cos\eta/c^2}\,, \\ w_2 &= \frac{v\sin\eta}{\alpha(1 - uv\cos\eta/c^2)}\,, \\ w_3 &= 0\,. \end{aligned}$$

Hence, from the point of view of O, the speed of O'' is

$$\begin{aligned} w &= \sqrt{w_1^2 + w_2^2} \\ &= \frac{\sqrt{(u - v\cos\eta)^2 + v^2\sin^2\eta/\alpha^2}}{1 - uv\cos\eta/c^2} \\ &= \frac{\sqrt{u^2 - 2uv\cos\eta + v^2 - u^2v^2\sin^2\eta/c^2}}{1 - uv\cos\eta/c^2}\,. \end{aligned}$$

It is easy to verify that w is less than c if u and v are each less than c. In fact (see Problem 1.9), there is an important relation that is satisfied by the three velocities:

$$\Bigl(1 - \frac{w^2}{c^2}\Bigr)\Bigl(1 - \frac{uv\cos\eta}{c^2}\Bigr)^2 = \Bigl(1 - \frac{u^2}{c^2}\Bigr)\Bigl(1 - \frac{v^2}{c^2}\Bigr)\,.$$

[17]We remark here that this angle is measured by O' using *Euclidean* measuring instruments, even though the resulting triangle will not have sides and angles that satisfy Euclidean relationships. There is no paradox here. An angle is a physically and geometrically dimensionless quantity representing an amount of rotation, and rotation is absolute in all three geometries: elliptic/spherical, parabolic (Euclidean), and hyperbolic. A right angle is one-fourth of a complete rotation. It forms a natural unit of angular measure, and is universally assigned the numerical value $\pi/2$. The reason for using that seemingly arbitrary value comes from analysis; Taylor series become very cumbersome if any other measure is assigned to a right angle. We shall call the resulting values of all angles simply the *numerical values* of those angles. We do not like to use the term *radian measure*, since it suggests what is true only in Euclidean geometry—that the numerical value of the central angle subtended by a circular arc is the ratio of the length of the arc to the length of the radius *and is independent of the radius.* In hyperbolic and elliptic geometry, the numerical value of the angle is not normally equal to that ratio, and the ratio itself varies with the radius of the circle.

[18]Since O is moving along O''s negative first spatial axis, the angle η must be measured *clockwise* from that axis to the line of motion that O' ascribes to O''. That accounts for the negative sign in the first coordinate of the location of O'''s origin, as seen by O'.

REMARK 1.5. This last equality relates the three magnification factors in the Lorentz transformations corresponding to the three velocities. That is, if $\alpha = c/\sqrt{c^2-u^2}$, $\beta = c/\sqrt{c^2-v^2}$, and $\gamma = c/\sqrt{c^2-w^2}$, then

$$\gamma = \alpha\beta\Big(1 - \frac{uv\cos\eta}{c^2}\Big)\,.$$

We have now established the simple but fundamental fact that if O'''s velocity relative to O' is constant in direction and constant (less than c) in magnitude, as judged by O', while O''s velocity relative to O is similarly constant in direction and constant (less than c) in magnitude relative to O, then O'''s velocity relative to O is also constant in direction and constant (less than c) in magnitude. Therefore the composition of two relativistic velocities is a relativistic velocity and hence a suitable (privileged) pair of coordinate systems that can be chosen by O and O'' should be related to each other by equations of the simple form derived above for the Lorentz transformation. Making that fact computable and verifying that the implied composition is associative if a fourth observer O''' comes along will occupy the last few sections of the present chapter. In the meantime, we wish to develop the trigonometry of these relativistic velocity triangles.

Recalling our remark above that if $u = c\tanh(U/k)$, then $\alpha = \cosh(U/k)$, we have, upon replacing β and γ by $\cosh(V/k)$ and $\cosh(W/k)$, the fundamental relation

$$\cosh(W/k) = \cosh(U/k)\cosh(V/k)(1 - \tanh(U/k)\tanh(V/k)\cos\eta)\,,$$

which, when the parentheses are removed, yields

$$\cosh(W/k) = \cosh(U/k)\cosh(V/k) - \sinh(U/k)\sinh(V/k)\cos\eta\,.$$

This last equality is precisely the law of cosines in a hyperbolic plane whose radius of curvature is $k\sqrt{-1}$ (see Appendix 1). Putting the matter another way, under the correspondence $u \leftrightarrow U$, where $u = c\,\tanh(U/k)$ and $U = k\,\ln\sqrt{(c+u)/(c-u)}$, the relativistic velocity triangle with sides u, v, w corresponds to a triangle with sides U, V, W, and if the latter is regarded as being in the hyperbolic, plane, then the two triangles have the same angles. Thus, at least as far as the law of cosines goes, the trigonometry of a relativistic velocity triangle is identical to the trigonometry of a triangle in a hyperbolic plane. Who could have guessed, two centuries ago, when this geometry was invented, that it would turn out to describe a world of physical laws that had not yet been imagined?

The line from O to O'' lies in the plane of O, O', and O'', and, as will be shown in the optional material that now follows, it makes an angle ξ with the positive x-axis, where

$$\begin{aligned}\cos\xi &= \frac{w_1}{w} = \frac{u - v\cos\eta}{w(1 - uv\cos\eta/c^2)} = \frac{u - v\cos\eta}{\sqrt{u^2 - 2uv\cos\eta + v^2 - u^2v^2\sin^2\eta/c^2}}\,,\\ \sin\xi &= \frac{w_2}{w} = \frac{v\sin\eta}{(w\alpha(1 - uv\cos\eta/c^2))}\\ &= \frac{v\sin\eta}{\alpha\sqrt{u^2 - 2uv\cos\eta + v^2 - u^2v^2\sin^2\eta/c^2}}\,.\end{aligned}$$

At this point, we have introduced the Lorentz transformation and discussed some of its interesting properties. That is all the background necessary to a discussion of the special theory of relativity, which begins in Chapter 2. The reader looking for a compact account of special relativity can therefore ignore the remaining sections of the present chapter, which have been written to explore certain matters of interest mostly to mathematicians.

9. Composition of Relativistic Velocities as a Binary Operation*

Consider once again our three observers O, O', and O'', any two of whom have a constant relative velocity. Assuming their origins all three coincide at some time, which each of them can take as the epoch of a calendar (time zero), any observer will see that they form a triangle at any fixed instant of time, and all observers will agree that the length of each side of the triangle is increasing after the initial instant when the three are at the same location. The trouble is, there is no instant except the time when all three origins coincided for which even two of the three observers O, O', and O'' can agree about the locations of all three of their origins, and different observers will observe different rates of increase for each of the sides. This difficulty raises a purely mathematical problem, which we may formulate as follows.

Each observer has a Euclidean space and a separate time line in which to record events $(t; x, y, z)$. In those coordinates O can define the velocity of O' as a vector $\boldsymbol{u}$, and similarly O' can define the velocity of O'' as a vector $\boldsymbol{v}$. Now O'' has some velocity, say $\boldsymbol{z}$ relative to O. How is $\boldsymbol{z}$ related to $\boldsymbol{u}$ and $\boldsymbol{v}$? Schematically, we can write this relation as a binary composition $\boldsymbol{z} = \boldsymbol{u} * \boldsymbol{v}$. In Newtonian mechanics, this binary operation is just the familiar vector addition in $\mathbb{R}^3$, and it is a commutative and associative operation. The symmetry of the formula for composition of relativistic velocities along a line shows that it is also commutative in that context. But the formulas we just gave for composition of nonparallel velocities show that in relativity it is generally not commutative. If we ignore certain difficulties, we can say that this binary operation is associative. That is, if we write $\boldsymbol{u}$ as the pair (O, O') and $\boldsymbol{v}$ as the pair (O', O''), then agree that, in general, if the second element of the first pair is the same as the first element of the second pair, we have the binary operation $(a, b) * (b, c) = (a, c)$. It is then obvious that $\big((a, b) * (b, c)\big) * (c, d) = (a, c) * (c, d) = (a, d) = (a, b) * \big((b, c) * (c, d)\big)$. But that simple computation already points up a seemingly insuperable difficulty. We just said that this composition is commutative when the velocities are collinear. But what can that mean in this notation? How can we compose (b, c) with (a, b)? We would need $c = a$ in order to do so. In terms of computation, what would it mean to assign velocity $\boldsymbol{v}$ to O' relative to O and $\boldsymbol{u}$ to O'' relative to O'?

Vectors have to be written as triples of real numbers, representing length and direction in someone's rectangular coordinate system. Whose system is this? We presume that $\boldsymbol{u}$ represents a vector in O's system and $\boldsymbol{v}$ a vector in O''s system. What does $\boldsymbol{u}$ mean in O''s system? We avoided this difficulty in our derivation of the Lorentz transformation by assuming that the two observers chose their axes so that the x-axis was along their common direction of motion. If we cut them free from this anchor, then we are computationally "adrift." What this means is that we have not yet pinned down the proper algebraic/geometric representation of relativistic velocities. We need some conventions that all of our observers can agree on as to the magnitude and direction of a velocity. Only after we get those conventions, which we shall do using hyperbolic plane trigonometry, will we be in a position to write the concatenation of two relativistic velocities in a form that computers can accept and work with.

There is a second difficulty as well, again connected with the special choice of axes used in deriving the Lorentz transformation. Because of the special choice of coordinates that we made, we got a simple "standard-form" matrix to represent the transformation. It is symmetric, and ten of its twelve nondiagonal entries are equal to zero. But this matrix takes account of only the relative speed of the motion; it assumes that the axes have already been adjusted to take into account the direction of that motion. If two velocities that we are composing are not parallel, multiplying the corresponding matrices does not yield the matrix of the composite velocity, since the simplified matrices relating O and O' to O'' have to be computed in axes that are rotated relative to those used for the matrix relating O to O'. After we solve the first problem, thereby making it possible to say in a computational sense that composition of relativistic velocities is an associative operation, we will be in a position compose the matrices of two standard-form Lorentz transformations and get the standard-form matrix of the composition of the two velocities that they correspond to. (As the reader may already have guessed, the procedure is to sandwich each matrix between two rotations of $\mathbb{R}^4$ representing the alignment of the first spatial axis of each observer with the lines of relative motion of the other two.)

9.1. Relativistic velocity triangles. With so much disagreement between different observers, we may well flee for refuge into the "safe" areas where there is agreement between different observers. What are these areas? Since the Lorentz transformation is linear, it does preserve lines and planes. Thus, each observer, "watching" the other two from the origin of his spatial coordinate system, will see them along two rays from his own origin. Any two of these observers will agree on their mutual speed. These lines do not change over time, and so the plane determined by the three observers is constant for each of them and may be regarded as common to all of them. We can then draw a triangle and label each of its vertices with a symbol representing one of the observers, assigning to each of its sides a "length" (velocity) agreed upon by the two observers at the endpoints. The key to the mutual communication we need for our observers is the fact that angles belong to absolute geometry: A right angle is one-fourth of a complete rotation, whether in elliptic, Euclidean, or hyperbolic geometry and is assigned the "radian" measure $\pi/2$ in all three versions of trigonometry, even though it has the interpretation as the ratio of the subtended arc of a circle to its radius only in Euclidean geometry.[19]

We have seen that two observers in relative motion do not generally agree about the angle between two lines. Nevertheless, each observer is "standing guard" at his own vertex of this triangle and can measure the two sides (velocities) adjacent to

[19]This is because circles have an absolute meaning in all three geometries, and arc length on a circle is rotation-invariant, just like the central angles subtended by arcs on the circle. On a given circle in any of these geometries, central angles and the arcs they subtend are proportional. The value $\pi/2$ for a right angle is based on the Euclidean case, where the measure of an angle really is the ratio of the arc it subtends to the radius. While this value is only a convention, it is extremely practical and will be called the *measure* of a right angle (not the *radian measure*, since this ratio is not the same for circles of different radii in elliptic and hyperbolic geometry).

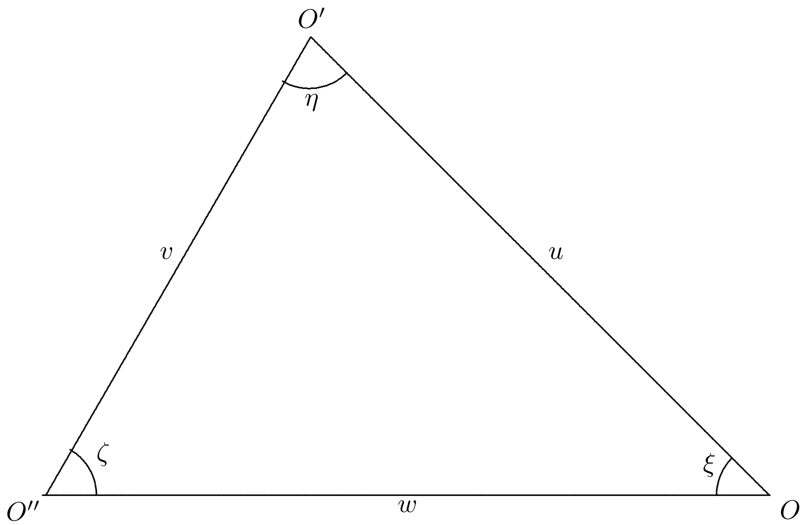

FIGURE 1.3. The velocity triangle

it and the angle between them. The angle measured by the observer located at each vertex is the angle between his lines of sight to the other two observers. Since the velocities are assumed constant, that angle does not change over time, and it is therefore taken as the definition of the angle of the velocity triangle at that vertex.

Each observer knows the two sides and included angle where he is located, and each pair of observers agrees about the magnitude of the side that each measures as the speed of the other. The set of three sides and three angles that results will be called a *relativistic velocity triangle*. If one of the observers computes the magnitude of the side opposite to his vertex using the Euclidean law of cosines, the result will *not* be the value that the other two observers agree on. That computed velocity might well be larger than c when computed in this way. That means only that two objects in an observer's space can move faster than light *relative to each other, as judged by the observer*, but not relative to the fixed frame of axes used by the observer. Since the sides of these triangles do not represent lengths, we need not think of them as situated in the Euclidean space familiar from geometry. The assignments of sides and angles that we have made are what three actual observers would measure in reality, however; and we can transfer them into a geometric space where the sides are lengths via the correspondence $u \leftrightarrow c\,\tanh(U/k)$, $U = k\,\mathrm{arctanh}\,(u/c)$ mentioned earlier, retaining the assignment of angles already made. They now form ordinary triangles in a three-dimensional space, but the geometry of that space is hyperbolic rather than Euclidean.

In Newtonian mechanics, an observer looking at the origins of the three systems at any given time would see them at the vertices of a physical triangle having the angles shown in the velocity triangle and sides proportional to those shown. The relations between the parts of the observed triangle will satisfy the Euclidean law of cosines:

$$w^2 = u^2 + v^2 - 2uv\cos\eta\,.$$

In relativistic mechanics, however, there is no instant at which any given observer would see a triangle having these angles and also having sides proportional to these sides, and so the triangle shown in Fig. 1.3, while it makes perfectly good sense in the space of velocities, is not similar to any triangle in the space of positions that would be observed by anybody. Its sides and angles have definite values according to the laws of special relativity, but these parts do not satisfy the Euclidean relations. In fact, the law of cosines in the relativistic velocity triangle is, as we have already calculated:

$$w^2 = \frac{u^2 + v^2 - 2uv\cos\eta - u^2v^2\sin^2\eta/c^2}{\bigl(1 - uv\cos\eta/c^2\bigr)^2}\,. \tag{1.5}$$

This law of cosines was derived in Einstein's 1905 paper on special relativity. Assuming that it holds for each angle in the triangle, it allows us to express the side opposite that angle in terms of the angle and the two sides adjacent to it.

The law of cosines is the basic rule for all trigonometry. Since a triangle is to be determined by any two sides and the included angle (Euclid's fundamental hypothesis about congruence of triangles), trigonometry faces the task of expressing all six parts of a triangle in terms of a given angle and its two adjacent sides. Obviously, it suffices to find an expression for the other two angles, since the third side is already determined by the law of cosines itself, and the two adjacent sides have to be given anyway. That task is achieved by the following formulas.

THEOREM 1.3. *The cosines and sines of angles ξ and ζ in Fig.* 1.3 *are given in terms of angle η and sides u and v by the following formulas:*

$$\cos\xi = \frac{u - v\cos\eta}{\sqrt{u^2 + v^2 - 2uv\cos\eta - u^2v^2\sin^2\eta/c^2}} = \frac{u - v\cos\eta}{w(1 - uv\cos\eta/c^2)}\,, \tag{1.6}$$

$$\sin\xi = \frac{v\sin\eta}{\alpha\sqrt{u^2 + v^2 - 2uv\cos\eta - u^2v^2\sin^2\eta/c^2}} = \frac{v\sin\eta}{\alpha w(1 - uv\cos\eta/c^2)}\,, \tag{1.7}$$

$$\cos\zeta = \frac{v - u\cos\eta}{\sqrt{u^2 + v^2 - 2uv\cos\eta - u^2v^2\sin^2\eta/c^2}} = \frac{v - u\cos\eta}{w(1 - uv\cos\eta/c^2)}\,, \tag{1.8}$$

$$\sin\zeta = \frac{u\sin\eta}{\beta\sqrt{u^2 + v^2 - 2uv\cos\eta - u^2v^2\sin^2\eta/c^2}} = \frac{u\sin\eta}{\beta w(1 - uv\cos\eta/c^2)}\,. \tag{1.9}$$

Here $\beta = c/\sqrt{c^2 - v^2}$. Notice that, when $\eta = \pi/2$ (the case of a right triangle) the expressions for $\cos\xi$ and $\cos\zeta$ agree with the Euclidean definitions, being u/w and v/w respectively. The sines are not the same as for a Euclidean triangle, being off by a factor of α for $\sin\xi$ and β for $\sin\zeta$.

PROOF. We take as our starting point the assumption that the relativistic law of cosines holds for each of the angles of a triangle, and we begin by showing how to derive Eq. (1.6) for the cosine of the angle ξ from this assumption. Applying the assumption to angles η and ξ yields

$$w^2 = \frac{u^2+v^2-2uv\cos\eta-(u^2v^2\sin^2\eta)/c^2}{\left(1-\frac{uv\cos\eta}{c^2}\right)^2}, \tag{1.10}$$

$$v^2 = \frac{u^2+w^2-2uw\cos\xi-(u^2w^2\sin^2\xi)/c^2}{\left(1-\frac{uw\cos\xi}{c^2}\right)^2}. \tag{1.11}$$

Using these equations to express $\cos\xi$ explicitly as a function of u, w, and η by eliminating v between the two equations leads to horrendously complicated algebraic expressions that even *Mathematica* cannot simplify. A simpler procedure is as follows: Let $x = \cos\xi$ and $y = \cos\eta$. Rewrite Eqs. (1.10) and (1.11) as

$$\begin{aligned} w^2 &= \frac{u^2+v^2-2uvy-u^2v^2(1-y^2)/c^2}{\left(1-\frac{uvy}{c^2}\right)^2}, \\ v^2 &= \frac{u^2+w^2-2uwx-u^2w^2(1-x^2)/c^2}{\left(1-\frac{uwx}{c^2}\right)^2}. \end{aligned}$$

This last equation is a quadratic equation in x:

$$v^2\left(1-\frac{uwx}{c^2}\right)^2 = u^2+w^2-2uwx-u^2w^2(1-x^2)/c^2. \tag{1.12}$$

This is a genuine quadratic equation, not degenerate, since the coefficient of x^2 is $-(u^2w^2)/(\beta^2c^2)$, where $\beta = c/\sqrt{c^2-v^2}$. This coefficient is not zero except in the trivial case when u or w is zero.

Next, eliminate all occurrences of w^2 by using the first equation. The first power w, however, should remain. The result is a new quadratic equation in x that can be solved to express x in terms of u, v, w, and y. This equation, however, is rather complicated, and a computer algebra system such as *Mathematica* is recommended to avoid algebraic mistakes here. One can then simply insert the value of x provided by Eq. (1.6), namely $x = (u-vy)/(w(1-uvy/c^2))$, and verify that the result is zero. Again, *Mathematica* makes this easier, though not trivially so; it must be coaxed by expanding and putting back together and simplifying before it yields the information that the equation is satisfied. It follows that the two applications of the relativistic law of cosines lead to an equation for $x = \cos\xi$ that is satisfied by two potential values of x, one of which is the value of $\cos\xi$ given by Eq. (1.6). To show that the other root could not be the cosine of any angle, one has only to observe that the constant term of the quadratic equation satisfied by x divided by the coefficient of x^2 is $c^4/(u^2w^2)$, which is larger than 1. Since this quotient is the product of the roots of the equation, the two roots cannot both lie in the interval $[-1,1]$, and therefore the other root is not the cosine of any angle. These computations are implemented in *Mathematica* Notebook 1 of Volume 3. $\square$

To sum up, we shall say that a *relativistic velocity triangle* with sides u, v, and w and opposite angles ζ, ξ, and η respectively, is one for which all three relativistic laws of cosines hold.

$$\begin{aligned} w^2 &= \frac{u^2 + v^2 - 2uv\cos\eta - u^2v^2\sin^2\eta/c^2}{\left(1 - \frac{uv\cos\eta}{c^2}\right)^2}, \\ v^2 &= \frac{u^2 + w^2 - 2uw\cos\xi - u^2w^2(\sin^2\xi)/c^2}{\left(1 - \frac{uw\cos\xi}{c^2}\right)^2)}, \\ u^2 &= \frac{v^2 + w^2 - 2vw\cos\zeta - v^2w^2(\sin^2\zeta)/c^2}{\left(1 - \frac{vw\cos\zeta}{c^2}\right)^2}. \end{aligned}$$

These laws imply the following relation between the cosines of the angles ξ and η opposite sides v and w respectively.

$$\cos\xi = \frac{u - v\cos\eta}{w\left(1 - \frac{uv\cos\eta}{c^2}\right)}. \tag{1.13}$$

All the relations among the velocities and angles measured by our three observers can be deduced from any three of them—which may even be the three angles ξ, η, and ζ—using these trigonometric formulas. What makes these triangles relativistic rather than Euclidean is the fact that each vertex is associated with an observer, whose *Euclidean/Newtonian* measurement of the two sides and angles at that vertex are used to assign measures to these parts of the triangle. Each pair of observers agrees about the velocity (length) of the side joining their vertices, but there is no such agreement about any of the angles. They are arbitrarily defined as the angles measured by their corresponding observers. In general none of the six parts of the triangle is agreed to by all three of the observers. Each observer, we emphasize, is using Euclidean geometry to measure the two sides and the angle at his vertex. That observer will consequently use the Euclidean law of cosines and Euclidean trigonometry to determine the opposite side and the other two angles, and each of the other two observers will generally disagree with him about all three of them. These values inferred from Euclidean geometry are therefore to be discounted when the observers reconcile their observations. By making use of the relativistic trigonometry just discussed, each observer can deduce the measurements the other two are making, and they can then agree on all their relative velocities.

Remark 1.6. As a final comment on the connection of relativistic velocity triangles with hyperbolic triangles, we shall now show that the formula for $\sin\xi$, when reinterpreted in the hyperbolic plane, is precisely the hyperbolic law of sines. It suffices to show that the formula for $\sin\xi$ translates under the given substitutions to become

$$\sin\xi = \frac{\sin\eta\sinh(V/k)}{\sinh(W/k)}.$$

By Equation (1.7), $u = c\tanh(U/k)$, $\alpha = \cosh(U/k)$, $v = c\tanh(V/k)$, $w = c\tanh(W/k)$, we have, upon taking account of the relativistic law of cosines,

$$\begin{aligned}\sin\xi &= \frac{\tanh(V/k)\sin\eta}{\tanh(W/k)\cosh(U/k)(1-\tanh(U/k)\tanh(V/k)\cos\eta)}\\ &= \frac{\sinh(V/k)\sin\eta\cosh(W/k)}{\sinh(W/k)\cosh(V/k)\cosh(U/k)(1-\tanh(U/k)\tanh(V/k)\cos\eta)}\\ &= \frac{\sinh(V/k)\sin\eta\cosh(W/k)}{\sinh(W/k)(\cosh(U/k)\cosh(V/k)-\sinh(U/k)\sinh(V/k)\cos\eta)}\\ &= \frac{\sinh(V/k)\sin\eta}{\sinh(W/k)},\end{aligned}$$

as required.

10. Plane Trigonometry*

Having established the values of ξ and ζ, we have now completed our earlier argument showing that the trigonometry of relativistic velocity triangles is simply hyperbolic trigonometry, that is, trigonometry on a sphere of imaginary radius $k\sqrt{-1}$ filtered through the mapping $u = c\tanh(U/k)$. Nevertheless, instead of constantly translating everything we want to say about velocities into statements in the hyperbolic plane, we prefer to deal directly with the formulas derived from the Lorentz transformation. We are not going to derive any facts about relativistic velocity triangles using hyperbolic trigonometry. We do not assume that the reader even knows hyperbolic trigonometry. Everything we prove geometrically is based on ordinary algebra and the Lorentz transformation. In fact, we do the exact opposite in Appendix 1: We use the Lorentz transformation as the key to a very simple treatment of the trigonometry of the hyperbolic plane, a task that is normally requires considerably more machinery. That is all we are going to say about this fascinating connection. For more details about non-Euclidean geometries, see Appendix 1.

In a simpler form, the expression for ξ is

$$\xi = \operatorname{arccot}\Big(\frac{\alpha(u-v\cos\eta)}{v\sin\eta}\Big) = \operatorname{arccot}\left(\alpha\Big(\frac{u}{v}\csc\eta-\cot\eta\Big)\right). \tag{1.14}$$

The angle ξ here is not a signed angle, since we are not at this point concerned with rotating O's coordinate system. It is simply an angle whose universal measure lies between 0 and π. (Again, we are deliberately not calling this universal measure *radian measure*.) The angle ζ that O'' observes between the lines joining his origin to the origins of O and O' can be obtained by interchanging u and v (and hence replacing α with β), and satisfies Eqs. (1.8) and (1.9). Again, expressed more simply

$$\zeta = \operatorname{arccot}\Big(\frac{\beta(v-u\cos\eta)}{u\sin\eta}\Big) = \operatorname{arccot}\left(\beta\Big(\frac{v}{u}\csc\eta-\cot\eta\Big)\right). \tag{1.15}$$

The law of cosines enables us to find the third side of a triangle given two sides and the included angle (see Fig. 1.3). The triangle has six parts—three sides and three angles—but only three of them are needed to determine the triangle. Let us briefly sketch how to infer the other three parts from three that are given.

(1) To solve a triangle given two angles and a side, we use the law of sines given in Problem 1.10 below. This law for relativistic velocity triangles is equivalent to the hyperbolic law of sines stated above.
(2) To find the angles given all three sides, see Problem 1.9 below.
(3) To solve a triangle given only its three angles ξ, η, and ζ opposite sides v, w, and u respectively, we combine the algebraic equations for the cosines of the angles (*Mathematica* is highly recommended for this exercise), and get the following formulas, which hold provided the expressions under the radicals are positive, as they must be if the sum of the three angles is less than two right angles.

$$\begin{aligned} u &= \frac{c\sqrt{\cos^2\xi + \cos^2\eta + \cos^2\zeta + 2\cos\xi\cos\eta\cos\zeta - 1}}{\cos\zeta + \cos\xi\cos\eta}, \\ v &= \frac{c\sqrt{\cos^2\xi + \cos^2\eta + \cos^2\zeta + 2\cos\xi\cos\eta\cos\zeta - 1}}{\cos\xi + \cos\eta\cos\zeta}, \\ w &= \frac{c\sqrt{\cos^2\xi + \cos^2\eta + \cos^2\zeta + 2\cos\xi\cos\eta\cos\zeta - 1}}{\cos\eta + \cos\zeta\cos\xi}. \end{aligned}$$

It is easy to verify that u, v, and w, as determined from these equations, are all smaller than c, provided all the denominators are positive, as must be the case if ξ, η, and ζ are to be the angles of a relativistic velocity triangle. In fact, any three positive angles ζ, ξ, and η, the sum of whose measures is less than π form the angles of a unique relativistic velocity triangle whose respective opposite sides are given by the three formulas above. The condition $\zeta + \xi + \eta < \pi$ guarantees that the expression under the radical and all three denominators are positive. (See Problem 1.17.)

10.1. Right triangles. When O and O'' are moving along mutually perpendicular lines as judged by O', these formulas naturally simplify. Taking $\eta = \pi/2$, we get the following equalities:

$$w = \sqrt{u^2 + v^2 - u^2v^2/c^2} = \frac{c}{\alpha\beta}\sqrt{\alpha^2\beta^2 - 1}\,, \tag{1.16}$$

$$\cos\xi = \frac{u}{\sqrt{u^2 + v^2 - u^2v^2/c^2}} = \frac{u\alpha\beta}{c\sqrt{\alpha^2\beta^2 - 1}}\,, \tag{1.17}$$

$$\sin\xi = \frac{v}{\alpha\sqrt{u^2 + v^2 - u^2v^2/c^2}} = \frac{v\beta}{c\sqrt{\alpha^2\beta^2 - 1}}\,, \tag{1.18}$$

$$\cos\zeta = \frac{v}{\sqrt{u^2 + v^2 - u^2v^2/c^2}} = \frac{v\alpha\beta}{c\sqrt{\alpha^2\beta^2 - 1}}\,, \tag{1.19}$$

$$\sin\zeta = \frac{u}{\beta\sqrt{u^2 + v^2 - u^2v^2/c^2}} = \frac{u\alpha}{c\sqrt{\alpha^2\beta^2 - 1}}\,. \tag{1.20}$$

EXAMPLE 1.1. The vector addition of velocities and displacements was discussed by Galileo in the seventeenth century. To use a modification of his example, suppose a ship is moving at a rate of 12 km per hour relative to the Earth. If you walk from the port side to the starboard side at 5 km per hour, then your actual speed is $\sqrt{12^2 + 5^2} = 13$ km per hour, in a direction that makes an angle $\arctan(5/12) \approx 22.6°$ with the direction the ship is moving. Relativity changes this relation. Since the speed of light is about $300,000$ km per second, if Observer O sees

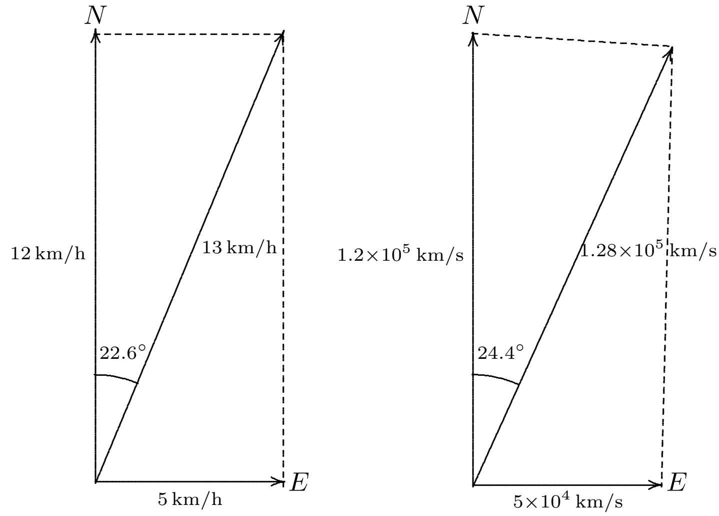

FIGURE 1.4. Galilean addition of velocities (left): A land-based observer sees a ship moving through the water heading due north at 12 km per hour. The pilot of the ship sees a crew member walking from port to starboard (due east) at 5 km per hour. The land-based observer sees that crew member moving at 13 km per hour in a direction 22.6° east of north. Relativistic addition of velocities (right): Three "observers" O, O', and O'' all move through the same origin simultaneously, then diverge. Observer O sees O' moving due north at 120,000 km per second, and O' sees O'' moving due east at 50,000 km per second. Then O sees O'' moving at 128,452 km per second in a direction 24.4° east of north.

observer O' moving at a speed of 120,000 km per second, and O' sees O'' moving at a speed of 50,000 km per second in a direction perpendicular to the direction in which he sees O moving, then O will see O'' moving at a speed of 128,452 km per second in a direction making an angle of 24.4° with with the line of sight to O'.

What can we make of a geometry in which u, v, and w are the sides of a right triangle, with u opposite angle ζ, v opposite angle ξ, and w opposite the right angle? In terms of the famous Pythagorean theorem, these relations would imply $u^2 + v^2 > w^2$. That is, the hypotenuse of a right triangle is shorter than the hypotenuse of a plane right triangle in Euclidean space having legs of the given lengths. That inequality is true of both elliptic non-Euclidean geometry, where the Pythagorean relation is $\cos(C/R) = \cos(A/R)\cos(B/R)$, and hyperbolic non-Euclidean geometry, where the corresponding Pythagorean relation is $\cosh(C/k) = \cosh(A/k)\cosh(B/k)$. (Here R and k are respectively the constant curvature of the elliptic plane and the hyperbolic plane. In elliptic geometry, the sum of the angles of a triangle is larger than two right angles, and in hyperbolic geometry it is less. We shall now show that the angle sum of a right relativistic velocity triangle is less than two right angles, that is, $\xi + \zeta < \pi/2$. We of course already know this fact, since it is a fact of hyperbolic geometry. Nevertheless, we shall give two direct proofs of it, one here for right triangles, and one later for general triangles.

THEOREM 1.4. *Let ξ and ζ be the two acute angles of a relativistic velocity triangle containing a right angle. Then $\xi + \zeta < \pi/2$.*

PROOF. By formulas (1.18) and (1.19), we have $\alpha \sin \xi = \cos \zeta = \sin(\pi/2 - \zeta)$. Since $\alpha = c/\sqrt{c^2 - u^2} > 1$, it follows that $\sin \xi < \sin(\pi/2 - \zeta)$, and therefore $\xi < \pi/2 - \zeta$, that is, $\xi + \zeta < \pi/2$, as asserted. □

It is not difficult to show (see Problem 1.13) that the sum of the three angles ξ, η, and ζ in any relativistic velocity triangle is less than two right angles.

In the special case when $u = v$, we get what O' regards as an isosceles right triangle, with acute angles ξ given by

$$\xi = \arccos\left(\frac{\alpha}{\sqrt{\alpha^2 + 1}}\right) = \arccos\left(\frac{1}{\sqrt{1 + 1/\alpha^2}}\right).$$

Obviously, ξ is a decreasing function of α. The limiting cases are $u \uparrow c$, in which $\xi \downarrow 0$, and $u \downarrow 0$, in which $\xi \uparrow \pi/4$, the latter being the case when all observers are at rest relative to one another, so that the geometry becomes Euclidean, and simultaneity becomes observer-independent.

REMARK 1.7. Although in general three observers do not agree about the velocities in the triangle whose vertices they form, there is a range of vertex angles η in an *isosceles* triangle for which there is a speed u (depending on η) on each of the two sides of the angle η such that the observer O' at the vertex of the angle will agree with O and O'' as to their mutual speed. The range of angles η for which such a speed u exists is

$$60^\circ = \arccos(1/2) < \eta < \arccos(1/3) < 70.528779365509308630755^\circ\,.$$

Over this range, the corresponding speed u decreases from c to 0. (See Problem 1.11 below.) Observers O and O'' do not agree with O' about any of the angles, and neither of them thinks this is an isosceles triangle. (For example, O does not agree with O' and O'' that their relative speed is $v = u$.)

As an example, take $u = (3\sqrt{2}/5)c$ and $\cos(\eta) = 5/12$. With those values, you will find that

$$\eta \approx 67.056553501352011261^\circ\,.$$

For this case, all three observers agree as to the relative speed of O and O'', namely $0.84c$.

11. The Lorentz Group*

Two problems may be of interest to the mathematically-inclined reader—perhaps less interesting to nonmathematicians. The first is the question of closure: Is the composition of two Lorentz transformations truly a Lorentz transformation? Certainly, if C is moving with constant velocity relative to B and B is moving with constant velocity relative to A, then C is also moving with constant velocity relative to A. Since our relativistic velocity triangles tell us how A and C can rotate coordinates so as to get the matrix of this transformation in the standard form, this question has, in a sense, a trivial answer. But do the computations really work out? We take up that question in Section 12 below and exhibit its computational implementation with *Mathematica* Notebook 5 in Volume 3.

The second question is the associativity of the law of relativistic composition of velocities. Suppose we have four observers, say A, B, C, and D, and that the velocity of B relative to A is $\boldsymbol{u}$, that of C relative to B is $\boldsymbol{v}$, and that of D relative to C is $\boldsymbol{w}$. Then the velocity of C relative to A can be represented as $\boldsymbol{u} +_L \boldsymbol{v}$, where this notation indicates the composition of velocities given by the relativistic velocity triangle determined by A, B, and C, as above. And then, by our computations, the velocity of D relative to A must be $(\boldsymbol{u} +_L \boldsymbol{v}) +_L \boldsymbol{w}$. On the other hand, the velocity of D relative to B is $\boldsymbol{v} +_L \boldsymbol{w}$, and consequently—if we believe the computations that have been performed above, the velocity of D relative to A must also be $\boldsymbol{u} +_L (\boldsymbol{v} +_L \boldsymbol{w})$. Since these two velocities are obviously the same, and we don't think we have made any mistakes in logic during our derivations, we have to conclude that the binary operation $+_L$ is an associative operation:

$$(\boldsymbol{u} +_L \boldsymbol{v}) +_L \boldsymbol{w} = \boldsymbol{u} +_L (\boldsymbol{v} +_L \boldsymbol{w})\,.$$

This relation would appear to settle the question, but there are subtleties involved when we try to implement this rule computationally. The velocities here are given in three distinct coordinate systems, used by A, B, and C. In what sense can they be added at all? How can you add coordinates in one set of axes to coordinates in a completely different set? We can tame the problem a bit by passing to a three-dimensional hyperbolic space, in which case the associative law becomes a geometric theorem:

THEOREM 1.5. *If the sides and angles of three faces of a tetrahedron satisfy the relations of relativistic velocity triangles, then* (1) *the sides and angles of the fourth face are determined and* (2) *the fourth face is also a relativistic velocity triangle.*

The proof of this theorem forms Section 5 of Appendix 1 in Volume 2. The fact that Lorentz transformations are closed under composition and that their composition is an associative operation makes them into a group, called the *Lorentz group.* It is a six-dimensional Lie group.

11.1. Associativity*. The fact that the fourth face of the velocity tetrahedron is a relativistic velocity triangle when the other three are amounts to the *associative law* for the relativistic composition of velocities. It is quite obvious that velocity $\mathbf{0}$ is an identity for this composition, and *it seems* that the inverse of velocity $\boldsymbol{u}$ should be $-\boldsymbol{u}$. And indeed it is, when $-\boldsymbol{u}$ is *defined* as the velocity of O relative to O' given that $\boldsymbol{u}$ is the velocity of O' relative to O. As we have already pointed out above, the associativity of this *verbally described* operation is completely obvious, even without any computations. Thus we have turned the set of physically possible relative velocities in three-dimensional space into a group. We emphasize, however, that both the elements of this group and the group operation have so far only verbal descriptions. Even though we think of these velocities as vectors, they do not transform between observers the way vectors do.

Most importantly, although it is trivial that composition of mappings is an associative operation, there is a subtlety involved in the present case: Before two observers apply the Lorentz transformation as we defined it (using a 4×4 matrix) to convert their space-time coordinates, they must both, in general, perform a rotation of their spatial axes. The fact that the actual transformation is a "sandwich" consisting of two rotation matrices with a standard-form Lorentz matrix between them makes it far from obvious that the composition is associative. Our reasons for believing that it is, at this point, are physical and geometric. It would be desirable to have an algebraic proof of the fact. That is difficult to do on the basis of our definition of a Lorentz transformation; we give an "empirical" verification of it in Theorem 1.6 below.

To get a *computable* associative operation out of the relativistic addition of velocities, the triples we need are not the components of velocity in some particular frame of reference, but rather the relative speed of two observers and the polar angles along which the two observe each other relative to fixed frames they are using. Using what we have proved by means of the velocity tetrahedron, we are now in a position to state the associative law formally and verify that it "computes" as it should. We still will not quite have made the velocities into a group, even when we do that, due to the singularity of the polar coordinate system at the origin. But at least we can allow our three observers to use whatever coordinates they like, and as long as none of the velocities is zero, we can say exactly how each needs to rotate its axes in order to communicate with the other two using a vector formula. We are not going to give a formal proof of the procedure, however, but rather rely on an "empirical" proof, using *Mathematica* to generate random data and verify the associativity of the composition. (The formal verification takes too long, even for *Mathematica*.)

Since we wish to discuss the composition without invoking a privileged coordinate system, we shall make a "sandwich" out of the relative velocity of two observers O and O', writing it as (θ, u, φ), where θ is the angle measured counterclockwise from O's first coordinate axis to the line of motion O observes O' to be traversing, u is the speed of that motion, and φ is the angle measured counterclockwise from O''s first coordinate axis to the line of motion O' observes O to be traversing. Similarly, let (χ, v, ψ) represent the relative velocity of O' and O''. The speeds u and v are positive numbers between 0 and c and the four angles are any real numbers, equality being taken modulo 2π. We wish to define a binary operation "$+_L$" that we shall call *Lorentz addition* for these two triples of real numbers.

Remark 1.8. Before we discuss how to implement our velocity triangle as a binary operation and show that it is associative, we note that we appear to be defining the set of relativistic velocities in a plane as a *three-dimensional* object. As we have described it, the set of relativistic velocities in a plane could be represented geometrically as the product of two circles (represented by the two angles θ and φ) and the interval $(0, c)$ (representing the speed). You could picture it as a solid torus (anchor ring) in Euclidean space with its outer surface stripped off and the circle through the middle of its interior, representing speed 0, also removed. Topologically, the tricky part of making this object into a computable group is attaching that all-important identity corresponding to speed 0. When two of these triples happen to be opposites of each other, that is, $O'' = O$, $\varphi = \chi$, $u = v$, and $\theta = \psi$, the operation we are going to define does give 0 as the speed, but it also attaches two angles to that speed. That makes no sense, given that the origin of a polar coordinate system has no definable polar angle.

These velocities, one would think, are really two-dimensional objects. To make them two-dimensional, we need some transformation of pairs of angles (θ, φ) that

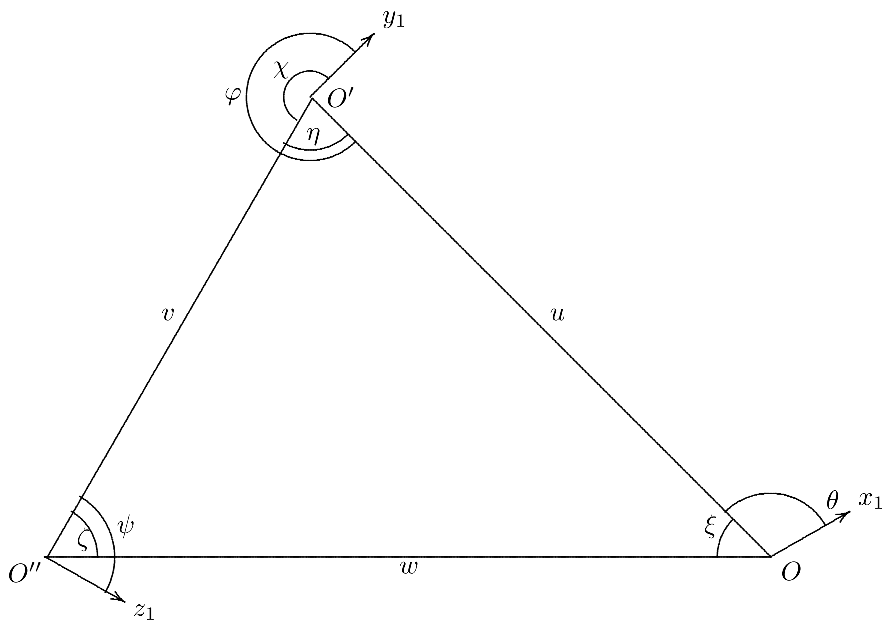

FIGURE 1.5. Lorentz addition of velocities

leaves the Lorentz sum of two velocities invariant. One transformation having this property is the mapping $(\theta, \varphi) \mapsto (d + \theta, d + \varphi)$ for any fixed angle d (see Problem 1.18). Once we have the appropriate transformation, we can take a "quotient space" modulo it and get the two-dimensional object that we need, except for that troublesome problem with the identity.

Defining the addition that we need is not complicated, given that we know how to solve the relativistic velocity triangle in Fig. 1.5

DEFINITION 1.1. The *Lorentz composition* $+_L$ of the speeds u and v in the directions indicated in Fig. (1.5) is the speed w in the direction shown, where we "sandwich" each speed between the two angles from the observer's first spatial axis in the directions of the other two observers. As a formula, replacing the vector velocity $\boldsymbol{u}$ by the triple (θ, u, φ), and similarly for $\boldsymbol{v}$ and $\boldsymbol{w}$, we get

$$(\theta, u, \varphi) \ \ +_L \ \ (\chi, v, \psi) = (\theta + \xi, w, \psi - \zeta)\,.$$

Since we have formulas for ξ, w, and ζ, the associativity of this operation ought to be straightforward.

We state it as a formal proposition:

THEOREM 1.6. *The composition* $(\theta, u, \varphi) \ \ +_L \ \ (\chi, v, \psi)$ *is an associative operation.*

PROOF. There is a complication resulting from the orientation of the triangle. If the angle $\eta = \varphi - \chi$ is larger than a straight angle, that is, $|\varphi - \chi| > \pi$, then the roles of O and O'' will interchange, and instead of $(\theta + \xi, w, \psi - \zeta)$, we would need $(\theta - \xi, w, \psi + \zeta)$. That will also happen if $|\varphi - \chi|$ is less than π but the y_1-axis is inside the angle η. (Since we represent angles as numbers between 0 and 2π, the angle χ will be larger than φ in this case.) To handle this complication, we need to multiply ξ and ζ by $\operatorname{sgn}(\pi - |\varphi - \chi|)$. Once that is done, although the computations are very tedious—so tedious that *Mathematica* will probably run out of memory before it can actually compute the Lorentz sum if the data are given as infinitely precise real numbers—it is possible to demonstrate convincingly through numerical examples that this operation is indeed associative. *Mathematica* Notebook 3 of Volume 3 will provide that convincing proof by generating as many random inputs as one likes. In that notebook, if you input two angles and a speed (the latter in the form ac, where a is a real number in the range $[0, 1)$) for each of two triples, the addition "$+_L$" (called `ladd` in *Mathematica* Notebook 3), can compute the Lorentz sum of the two relative velocities these triples represent, with the caveat that infinitely precise real numbers as data may well lead to a long computation requiring an inordinate amount of computer memory. If you input the data as finite-precision floating-point numbers (again, the speeds must bear the letter c as a suffix, even if they are zero), *Mathematica* will perform the computation in short order. The last command in the notebook checks the associative property by generating five triples of triples representing relative velocities and showing that the composition of three such velocities is the same, no matter how they are grouped.

In dozens of trials, this program always wrote
"`Out[4] = {{0,0,0},{0,0,0},{0,0,0},{0,0,0},{0,0,0}}`" .
Thus, by empirical verification, the operation `ladd` is associative. □

Although the theoretical basis of this program does not include the case of speed 0, the program will accept an input of $0.0c$ for either speed, and the output will be the other term in the sum, except that the two angles in it will be reduced modulo 2π.

12. Closure of Lorentz Transformations under Composition*

In the present section, we shall be considering a number of linear operators on space-time. To keep the algebra simple, it is desirable that the entries in these matrices be physically dimensionless. For that reason, we are going to make use of the technique mentioned earlier, "spatializing" the time coordinate through multiplication by the speed of light. Thus we shall replace time t by the variable $\tau = ct$. Also, to avoid getting too many primes, we shall rename our observers O, O' and O'', referring to them henceforth as X, Y, and Z. Finally, to streamline the notation still further, we assume these three observers are using the space-time coordinates $(\rho; \boldsymbol{\xi}) = (\rho; x^1, x^2, x^3)$, $(\sigma; \boldsymbol{\eta}) = (\sigma; y^1, y^2, y^3)$, and $(\tau; \boldsymbol{\zeta}) = (\tau; z^1, z^2, z^3)$ respectively. As always, we assume that the origins of all three four-dimensional coordinate systems coincide.

We have not actually given a *mathematical* definition of what a Lorentz transformation of $\mathbb{R}^4$ is. Our approach has been through kinematics: It is the transformation of space-time coordinates between two observers in uniform relative motion. What we have derived through simple algebra and Euclidean trigonometry is that, if both observers are using (Euclidean) orthonormal coordinates in $\mathbb{R}^4$ for which (1) the origins coincide, (2) the first axis represents time for both observers, and (3) the second axis is along the line of mutual motion for both observers, this transformation has a matrix of the form

$$L \sim \begin{pmatrix} \alpha & -\frac{\alpha u}{c} & 0 & 0 \\ -\frac{\alpha u}{c} & \alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where u is the speed of the second observer relative to the first and $\alpha = (1 - u^2/c^2)^{-1/2}$. More generally, since each observer can perform an orthonormal transformation on the spatial part of $\mathbb{R}^4$, a Lorentz transformation is one whose matrix M is such that there exist two rotation matrices R_1 and R_2 on the spatial portion of $\mathbb{R}^4$ (leaving the time axis fixed) such that $L = R_1 M R_2$ has this form. Or, more practically, we can define a Lorentz transformation to be the *set* of all matrices of this form, where M is a fixed matrix in standard form and R_1 and R_2 vary over all rotations of the spatial portion of space-time. When we use this description as a definition, it is by no means obvious that the Lorentz transformations are even closed under the operation of composition.

Mathematicians generally define a Lorentz transformation on $\mathbb{R}^4$ to be a linear transformation that preserves the quadratic form $\rho^2 - (x^1)^2 - (x^2)^2 - (x^3)^2$. In that form, it *is* obvious that the composition of two Lorentz transformations is a Lorentz transformation. Those concerned with logical rigor are then faced with two choices: (1) Demonstrating that the transformations fitting the description just given are precisely the ones that preserve the given quadratic form; (2) showing directly that if $L_1 = R_{11} M_1 R_{12}$ and $L_2 = R_{21} M_2 R_{22}$ are both Lorentz transformations, then there are rotations S_1 and S_2 of the given form such that $S_1 L_2 L_1 S_2$ has the given form. That is the route we shall follow, using *Mathematica* to avoid having to do messy computations by hand.

We now take up the computational problem just posed: getting the standard-form matrix of the composition of two velocities knowing relative speeds and lines of sight between the pairs of observers. The "sandwich" made by putting the two transformations between two rotations of a special form, which we created to compose relativistic velocities, makes this process computable, though a bit messy.

In our derivation of the Lorentz transformation between X and Y (O and O', as we called them at the time), we assumed that that the x^1- and y^1-axes coincide at all times, that Y is moving in the positive direction with speed u along this axis, as observed by X, without any rotation. This last means that $x^2 = y^2$ and $x^3 = y^3$ for any event E having coordinates $(\rho, x^1, x^2, x^3) = (\rho, \boldsymbol{\xi})$ and $(\sigma; y^1, y^2, y^3) = (\sigma; \boldsymbol{\eta})$, as measured by X and Y respectively. The FitzGerald–Lorentz contraction factor for this transformation is $1/\alpha$, where $\alpha = c/\sqrt{c^2 - u^2}$. In the three-dimensional velocity space used in common by X and Y, the velocity is $\boldsymbol{u} = (u, 0, 0)$, when both observers use the axes just described.

In these two coordinate systems, the Lorentz transformation corresponding to the velocity $\boldsymbol{u}$ can be written as the matrix equation

$$\begin{pmatrix} \sigma \\ y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} \alpha & -\alpha u/c & 0 & 0 \\ -\alpha u/c & \alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \rho \\ x^1 \\ x^2 \\ x^3 \end{pmatrix},$$

or equivalently, the set of four equations

$$\begin{aligned}\sigma &= \alpha(\rho - ux^1/c)\,,\\ y^1 &= \alpha(-u\rho/c + x^1)\,,\\ y^2 &= x^2\,,\\ y^3 &= x^3\,.\end{aligned}$$

REMARK 1.9. The matrix of the Lorentz transformation in these coordinates is obviously symmetric and hence can be diagonalized in $\mathbb{R}^4$ by a simple rotation of the ρx^1-plane (which is also the σy^1-plane) through the angle $\pi/4$ (half of a right angle). Its eigenvalues are 1, 1, $\alpha(1+u/c) = \sqrt{\frac{c+u}{c-u}}$, and $\alpha(1-u/c) = \sqrt{\frac{c-u}{c+u}}$. We remark that since all the eigenvalues are positive, this matrix is positive-definite, that is, it can be used to define a positive-definite quadratic form on $\mathbb{R}^4$. What is more important, however, is that this matrix preserves the space-time interval $\rho^2 - (x^1)^2 - (x^2)^2 - (x^3)^2$. It is easy to verify that this form is preserved, that is,

$$\rho^2 - (x^1)^2 - (x^2)^2 - (x^3)^2 = \sigma^2 - (y^1)^2 - (y^2)^2 - (y^3)^2\,.$$

The matrix of the inverse of the matrix of the Lorentz transformation in these coordinates is obtained by replacing u with $-u$, as we would expect.

Lorentz transformations are not the only linear transformations that preserve the space-time interval. An obvious group of linear transformations with this property corresponds to the set of matrices of the form

$$\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & r_{11} & r_{12} & r_{13}\\ 0 & r_{21} & r_{22} & r_{23}\\ 0 & r_{31} & r_{32} & r_{33}\end{pmatrix},$$

where the r's are the entries in an orthogonal matrix.

12.1. The standard-form matrix of the composite velocity. Suppose now that there is a third observer Z moving relative to Y with velocity $\boldsymbol{v}$, again without any rotation, but that $\boldsymbol{v}$ is not (necessarily) along Y's positive first spatial axis. Suppose it makes angle η with that axis. Then $\boldsymbol{v}$ makes angle $\theta = \pi - \eta$ with $-\,\boldsymbol{u}$ (the velocity that X has relative to Y). The angle θ is the angle between the lines of sight from Y to X and Z.

As we have seen, the vector representing the velocity of X relative to Z is generally not the negative of the vector representing the velocity of Z relative to X. The discrepancy is due to the changes of coordinates that X and Y must make when shifting from communicating with each other to communicating with Z. Each of the three observers needs to use two sets of coordinates: X will use coordinates $(\rho;\boldsymbol{\xi}') = (\rho;(x')^1,(x')^2,(x')^3)$ when communicating with Z and $(\rho,\boldsymbol{\xi})$ when communicating with Y; Y will use $(\sigma;\boldsymbol{\eta})$ when communicating with X and $(\sigma;\boldsymbol{\eta}') = (\sigma;(y')^1,(y')^2,(y')^3)$, when communicating with Z; and Z will use $(\tau;\boldsymbol{\zeta})$ when communicating with Y and $(\tau,\boldsymbol{\zeta}') = (\tau;(z')^1,(z')^2,(z')^3)$ when communicating with X. We plan to show that if X and Z communicate using the coordinates $(\rho;\boldsymbol{\xi}')$ and $(\tau;\boldsymbol{\zeta}')$, respectively, then the mapping $(\rho;\boldsymbol{\xi}') \mapsto (\tau;\boldsymbol{\zeta}')$ is given in matrix form as

$$\begin{pmatrix} \tau \\ (z')^1 \\ (z')^2 \\ (z')^3 \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma w/c & 0 & 0 \\ -\gamma w/c & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \rho \\ (x')^1 \\ (x')^2 \\ (x')^3 \end{pmatrix},$$

where $\gamma = c/\sqrt{c^2 - w^2}$, and the transformation is equivalently described by the set of four equations

$$\begin{aligned} \tau &= \gamma(\rho - w(x')^1/c)\,, \\ (z')^1 &= \gamma(-w\rho/c + (x')^1)\,, \\ (z')^2 &= (x')^2\,, \\ (z')^3 &= (x')^3\,. \end{aligned}$$

These relations are the result of composing the two Lorentz transformations with the rotations in the correct sequence. In particular, the 4×4 matrix of the composite transformation in the proper coordinate systems for O and O'' is the product of five 4×4 matrices, whose entries are rational functions of the entries in the two Lorentz matrices and the sines and cosines of the angles of rotation. The full computation is intimidatingly complicated, but a judicious use of *Mathematica* will verify its correctness.

THEOREM 1.7. *If L_α is the standard-form matrix of the Lorentz transformation corresponding to velocity $\boldsymbol{u}$, and L_β the standard-form matrix of the Lorentz transformation corresponding to velocity $\boldsymbol{v}$, which is to say $\alpha = (1 - u^2/c^2)^{-1/2}$, $\beta = (1 - v^2/c^2)^{-1/2}$, the directions being as in Fig.* 1.5, *and*

$$L_\alpha = \begin{pmatrix} \alpha & -\frac{\alpha u}{c} & 0 & 0 \\ -\frac{\alpha u}{c} & \alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } L_\beta = \begin{pmatrix} \beta & -\frac{\beta v}{c} & 0 & 0 \\ -\frac{\beta v}{c} & \beta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

then the composition $\boldsymbol{u} +_L \boldsymbol{v}$ has the standard-form matrix

$$L_\gamma = R_{-\zeta} L_\beta R_\theta L_\alpha R_{-\xi},,$$

where R_ψ represents a rotation of $\mathbb{R}^4$ that leaves the time axis and the third spatial axis fixed, and on the plane of the other two axes is a rotation about the third spatial axis through angle ψ. Moreover,

$$\gamma = \alpha\,\beta\,(1 - uv\cos\eta/c^2) = \alpha\,\beta\,(1 + uv\cos\theta/c^2)\,,$$

where $\theta = \pi - \eta$ is the angle between the lines of sight from Y to X and Z.

PROOF. Since we are going to be computing the transformation between X and Z, we shall assume at the outset that both of these observers have, if necessary, rotated their first spatial axes in the direction of each other, X being in the negative direction from Z and Z in the positive direction from X. Consider now the X-coordinates $(\rho; x^1, x^2, x^3)$ of an event. If these are to be transmitted to Y using the standard-form matrix representation of L_α, they must first be multiplied by a rotation matrix of the x^1x^2-plane through angle $-\xi$, where ξ is the angle between the lines of sight from X to Y and Z, that is,[20] by the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\xi & \sin\xi & 0 \\ 0 & -\sin\xi & \cos\xi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In order to receive this transmission through the standard-form matrix L_α, Y must use a coordinate system such that X is moving in the negative direction on Y's first spatial axis. Again, to simplify things, we assume that such is the case. The result of this transmission is the 4×1 matrix whose entries are the components of the vector (σ, y^1, y^2, y^3). Then, so that Y and Z can translate their coordinates with the simple 4×4 matrix corresponding to velocity $\boldsymbol{v}$, this 4×1 matrix needs to transformed via a rotation in the y^1y^2-plane through angle $\eta = \pi - \theta$, that is, it needs to be multiplied by the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The result of this multiplication is the matrix corresponding to $(\sigma; (y')^1, (y')^2, (y')^3)$, through which Y can communicate with Z via the Lorentz transformation

$$\begin{pmatrix} \tau \\ z^1 \\ z^2 \\ z^3 \end{pmatrix} = \begin{pmatrix} \beta & -\beta v/c & 0 & 0 \\ -\beta v/c & \beta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma \\ (y')^1 \\ (y')^2 \\ (y')^3 \end{pmatrix}.$$

Finally, Z must rotate the z^1z^2-plane through the angle $-\zeta$ in order to get the coordinates $(\tau; (z')^1, (z')^2, (z')^3)$ needed to communicate with O. That is, the 4×1 matrix corresponding to $(\tau; \boldsymbol{\zeta})$ is to be multiplied by

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\zeta & \sin\zeta & 0 \\ 0 & -\sin\zeta & \cos\zeta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

When all five of these matrices have been multiplied in the proper order, we do indeed find that the matrix of the mapping $(\rho; \boldsymbol{\xi}') \mapsto (\tau; \boldsymbol{\zeta}')$ is

$$\begin{pmatrix} \gamma & -\gamma w/c & 0 & 0 \\ -\gamma w/c & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The algebraic operations involved in verifying this fact are intimidatingly complex, and *Mathematica* should probably be invoked to shorten the labor. Before verifying that the computation is correct, however, we shall look at a few numerical examples. Again, a *Mathematica* program makes the work much easier, and we shall first present a simple one-cell *Mathematica* notebook (Notebook 4 in Volume 3) that shows how to get the five matrices we need and produces their product instantly. In principle, this program ought to be able to verify that the matrix

[20] We previously said X needed to rotate through angle ξ. But that was assuming X started with his principal axis "aimed" at Y. In the present discussion, X begins with his principal axis "aimed" at Z.

product we are discussing is the correct one. The algebra, however, is complicated, and *Mathematica* doesn't seem to be able to simplify it. As a result, we need to go interactive with *Mathematica*, using a new approach. *Mathematica* Notebook 5 in Volume 3 achieves this end. □

13. Rotational Motion and a Non-Euclidean Geometry*

The ratio of the earth to the heavens is that of a point...

There are some who disagree [with the claim that the Earth does not move] and propose to the contrary, thinking that there is no evidence to contradict them, that the heavens are fixed at a certain distance and that the Earth rotates about the same axis from west to east, making approximately one rotation every day, or that the two rotations [of the Earth and the heavens] can be combined in arbitrary amounts, provided only that they are about the same axis, as we said, and meet each other symmetrically.

It has escaped the attention of these people that, while the simpler mechanics of the stars are equally well explained under this assumption, when considered in relation to the properties that would follow from it as regards the air in which we ourselves live, it is seen to be utterly ridiculous.

Claudius Ptolemy (fl. ca. 130 CE), *Syntaxis* (*The Almagest*), Book 1, §§ 6, 7. My translation.

Comment: Several centuries earlier, Archimedes, in his *Sand-reckoner*, in which he calculated the number of grains of sand required to fill up the universe, pointed out that the first statement—obviously made long before Ptolemy lived—was absurd. Archimedes was applying the Euclidean definition of proportion, in which two objects can have a ratio only if they are of the same nature and some multiple of each is greater than the other. No multiple of a point can be larger than a line, but Ptolemy was a practical man, and the image of the ratio of a point to a line is useful in conveying his ideas.

Ptolemy's fear that the earth would disintegrate [if it moved], or that any natural motion—which acts very differently from an artificial or human-caused motion—could cause it to disintegrate, was therefore groundless. He should rather have worried about the universe, which must be moving much faster, being so far away from the earth...

If this ratio [of a point to a line, which Ptolemy gave as the ratio of the Earth to the heavens] held in fact, then the heavens would be infinitely far away... The farther away they are, the faster they must move in order to make a complete circuit in 24 hours. The increasing distance and velocity would each cause the other to become infinite, and it is a principle of Natural Philosophy that what is infinite cannot undergo any change or motion. Therefore, the heavens are at rest.

Nicolaus Copernicus (1473–1543), *On the Revolutions of the Heavenly Spheres*, Book 1, § 8. My translation.

Comment: The reasoning seems to be that, since the Earth is *not* a point, anything having such a ratio to it would have to be infinitely large.

> *In Ptolemy's system, the motions of the heavenly bodies cannot be explained by the action of central forces; celestial mechanics is impossible. The intimate relations that celestial mechanics reveals to us between all the celestial phenomena are true relations; to affirm the immobility of the earth would be to deny these relations; that would be to fool ourselves.*
>
> Henri Poincaré ([**65**], Chapter XI, § VII, p. 352).

> *An observer at the common origin [of two planes mutually revolving about the same axis through that point] who is able to observe clocks located on a given circle with center at the origin would thus see that a rotating clock runs slower than a fixed clock adjacent to it. Since he will not thereby be led to admit that the speed of light along the given radius depends on time, he will interpret the observation by saying that the rotating clock "really does" run slower. He will thus not be able to avoid defining time in such a way that the rate at which a clock runs depends on its location.*
>
> Einstein ([**21**], p. 775). My translation.

> *Even on the most modern views, the question of absolute rotation presents difficulties. If all motion is relative, the difference between the hypothesis that the Earth rotates and the hypothesis that the heavens revolve is purely verbal; it is no more than the difference between "John is the father of James" and "James is the son of John." But if the heavens revolve, the stars move faster than light, which is considered impossible. It cannot be said that the modern answers to this difficulty are completely satisfying...*
>
> Bertrand Russell ([**70**], p. 540).

Many puzzles arise as a result of the relativity of space and time and the FitzGerald–Lorentz contraction. Among them are the car wash puzzle, explored in Problem 1.19 below, and the twin paradox already discussed (see also Problem 1.2 below).

A third puzzle, to which the present section is devoted, involves astronomy: If it truly makes no difference which of two objects is moving, in what sense was Copernicus right and Ptolemy wrong about the solar system? Whether the Earth rotates on an axis or the whole universe revolves around the Earth should make no difference. On the other hand, if the stars revolve around the Earth once a day, then each of them is traversing a gigantic circle many light-years in radius, yet doing so in the space of a single day. How can this seeming contradiction be explained? That is the puzzle we now attempt to solve. The example we are going to give was described qualitatively by Einstein in his 1916 paper on general relativity (see the quotation above). What we intend to do is give the mathematical details *from the standpoint of special relativity.* A full explanation requires the general theory, and one of the purposes of including this example is to make that point. We emphasize that it is the special theory that we are using for this analysis, and that it should not be taken seriously as astronomy. It represents what Eddington ([**16**], p. 113) called"a crude application of the FitzGerald formula." Using the general theory of relativity, Eddington showed that the actual contraction of the circumference was only one quarter of that obtained from the analysis we are about to give. Lorentz [**56**] had shown earlier that a rotating solid disk could not contract to the full amount, since elastic forces within it would resist the contraction. Since we are talking about a completely empty astronomical orbit, however, elastic forces would not appear to be involved; and since we are only imagining the scene anyway, it does no harm to use the principles of special relativity only.

In Newtonian terms, if we take Ptolemy's point of view, then the stars are not moving in straight lines at constant speed and therefore some force must be moving them. In the relativistic version of astronomy presented in Chapter 4 below, they are not moving along geodesics (see Appendix 2) even in the geometry we are going to create for Ptolemy's benefit. Thus, even if we manage to prevent Ptolemy's framework from becoming totally absurd, we still leave him "on the hook" for an explanation of the fact that material bodies are moving in circles without any forces in evidence to make them do so.[21] To say this is not to criticize Ptolemy. Neither he nor Copernicus nor Kepler, who refined the heliocentric model by introducing elliptical orbits, aimed at creating a dynamical model of the universe. Their models were purely kinematic, meant to "save the appearances," that is, to hypothesize motions for the heavenly bodies that explain what we see from the Earth. The idea of linking astronomy with the mechanics of forces had to await the recognition of the importance of acceleration. This concept slowly came into focus over a period of time from the thirteenth century on, beginning with the Merton rule for the distance covered by uniformly accelerated bodies, followed by Galileo's application of it to bodies falling near the Earth's surface. Soon after (1644), Descartes stated what we now recognize as the law of inertia and the law of conservation of momentum, all of which Newton wrapped up neatly in his 1687 masterpiece *The Mathematical Principles of Natural Philosophy.* This fundamental change in the way mechanics was treated began in the decade after the death of Kepler, a full century after the death of Copernicus and a millennium and a half after the time of Ptolemy.

We are going to show that the old Ptolemaic (geocentric) system of astronomy can be used to explain celestial *kinematics* (not its *dynamics*) just as completely as the Copernican (heliocentric) system. As for the relativistic dynamics of a rotating universe we shall discuss that subject very briefly in the context of general relativity (without giving any details), in Chapter 7. At that point, we shall have the Einstein tensor at our disposal, making it possible at least to state comprehensibly the fascinating exact solution of the Einstein field equations given by Gödel [**34**], in which time travel is possible!

In Ptolemy's system, the stars revolve around the Earth once a day. By Euclidean reckoning, light cannot traverse a circle more than 5 billion kilometers in radius in the course of a single day, and the nearest star is more than 30 trillion kilometers away. Therefore some non-Euclidean geometry will be required. To understand the problem, we briefly recap what was done in the case of translational motion, with the Lorentz transformation. The secret of reconciling the points of view of two observers in relative motion along a straight line is that, while each

---

[21] Ptolemy himself, of course, was not bothered by this fact. In his cosmology, circular motion was the natural motion of all bodies in the heavens, and no further explanation of it was needed.

is using separate time and spatial coordinates and the proper space of each is Euclidean, from the point of view of the other observer, those time and space coordinates are intertwined, and the space is, as a result, not Euclidean. The Lorentz transformation was derived for the case of rectilinear motion at constant speed, which classically requires no forces.

The length of the radius from the center of rotation to each point should be the same for both observers, since the motion of that radius is perpendicular to itself at every point, and therefore it does not undergo any FitzGerald–Lorentz contraction.[22] We can assume that our two observers, whom we shall call Ptolemy and Copernicus, share the same clock at the origin. We do not need to worry about the synchronization of clocks that each has situated at what he regards as fixed points on a circle about the origin. Each of them thinks the clocks belonging to the other are whizzing by his own clocks, and, as Einstein remarked, that means each of them thinks the other's clocks are running slow. The two must disagree as to the length of the orbit of a star. Copernicus says the stars are fixed, and therefore the circumference of a circle centered at the Earth and passing through a star at distance R is $2\pi R$. His geometry is, by his measurements, Euclidean. If we are going to tailor our theory to fit the facts in this case, we shall have to allow Ptolemy to shorten that circumference, making it less than the distance that light can travel in a single day. We must picture Ptolemy observing a star careering around its orbit and seeing the path it is going to travel as being shorter than Copernicus measures it.

That is the theoretical basis of what we are about to do. The rest is merely a matter of computational details, which we give below. As it turns out, Ptolemy can say that what Copernicus is regarding as a three-dimensional space partitioned into parallel flat planes is in fact a stack of surfaces of revolution resulting from curling up each of those planes in the same way, the entire stack forming the interior of an infinite cylinder whose axis is the axis of mutual rotation and whose radius is $C/2\pi$, where C is the distance that light can travel during the time of one complete revolution.

13.1. The computational details. We analyze this problem by imagining two observers at the center of the Earth, one (Ptolemy) treating the Earth as fixed, while the other (Copernicus) regards the stars as fixed. For purposes of discussion we focus attention on a single star, which Ptolemy thinks is revolving around him once a day, while Copernicus takes the view that it is Ptolemy who is rotating, the star remaining fixed. In this situation, Ptolemy and Copernicus agree about a number of things, including the distance R from the Earth to the star (since the revolution of each point is perpendicular to the radial coordinate to that point, hence produces no FitzGerald–Lorentz contraction of the radius). They also agree about the local time, since they both have the same coordinate origin at all times. Thus, they agree about the period T of revolution while disagreeing as to who is doing the revolving, and they agree as to the angular velocity of revolution $\theta' = 2\pi/T$.

[22]Due to the ambiguity in the concept of "now," however, a radial path through the stars observed at a given instant by one of the two observers would be a spiral as observed by the other. We don't worry about this fact, since we are confining our attention to a single circle with center at the origin.

Where they disagree is in the length C of the orbit of the star. For Copernicus, for whom the star is fixed in a Euclidean space, that length is simply $2\pi R$. Let us temporarily assign length C to the orbit in Ptolemy's scheme of things. The speed u of the star, according to Ptolemy, is $C/T = C\theta'/(2\pi)$. In his 1916 paper on general relativity, Einstein brought up the subject of a rotating frame of reference, and noted that the circumference of a circular orbit would undergo FitzGerald–Lorentz contraction. Developing that idea, we find

$$C = 2\pi R\sqrt{1 - C^2(\theta')^2/(4\pi^2c^2)}\,.$$

Solving this equation for C, we find

$$\begin{aligned} C &= \frac{2\pi R}{\sqrt{1+\frac{R^2(\theta')^2}{c^2}}} = cT\frac{2\pi R}{\sqrt{(2\pi R)^2+(cT)^2}}\,,\\ u &= \frac{C}{T} = \frac{2\pi Rc}{\sqrt{(2\pi R)^2+(cT)^2}}\,. \end{aligned}$$

These last expressions show that the relativistic length of the orbit C is smaller than cT (the distance light travels over a period of revolution), and that u is smaller than c, no matter how large R becomes.

As $c \to \infty$, or $R \to 0$, both of these expressions approximate the classical expressions for circumference $(2\pi R)$ and speed $(2\pi R/T)$. That is, for small R or large c, we have $cT/\sqrt{(2\pi R)^2 + (cT)^2} \approx 1$.

As $R \to \infty$, we find $C/cT = \frac{2\pi R}{\sqrt{(2\pi R)^2+(cT)^2}} \to 1$ and $u/c \to 1$.

We can represent the geometry of this relativistically rotating plane as the ordinary Euclidean geometry of a curved surface in three-dimensional space and at the same time make C the circumference of an actual circle in $\mathbb{R}^3$. Introducing the variable $r = C/2\pi$, we claim that such a surface has the following equation in cylindrical coordinates (r, θ, z):

$$z = z(r) = \frac{cT}{4\pi}\int_0^{(2\pi r)^2/(cT)^2} \sqrt{\frac{s^2-3s+3}{(1-s)^3}}\,ds\,,$$

where the domain is $r < \frac{cT}{2\pi}$. This surface is shown in Fig. 1.5. The radius of the orbit described by a point is R, and it is the length of a certain curve from the origin to the orbit. You can compute this length as

$$\int_0^r \sqrt{1+\Big(\frac{dz}{dt}\Big)^2}\,dt\,.$$

This integral is elementary, since its integrand is $(1 - (2\pi t/cT)^2)^{-3/2}$, and the integral works out to be

$$\frac{r}{\sqrt{1-(2\pi r/cT)^2}} = R\,.$$

The orbit of the star is a circular horizontal section of the surface having circumference $2\pi r$.

This example shows how a curved representation of physical space can be useful in physics. After we define curvature in Chapter 5, we will be able to verify that the curvature of this surface at radius r is

$$\kappa = 3\Big(\frac{2\pi}{cT}\Big)^2\Big(1-\Big(\frac{2\pi r}{cT}\Big)^2\Big)^2\,.$$

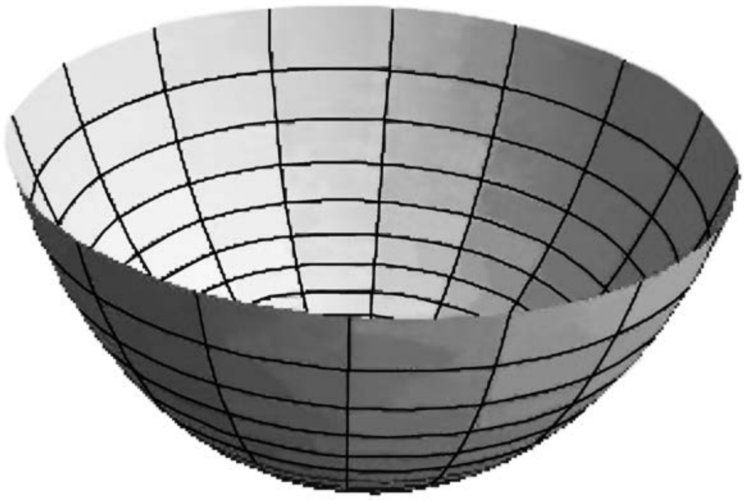

FIGURE 1.6. The relativistic geometry of a rotating plane

This expression shows that the curvature is small (the surface is nearly flat) if c is very large. After we discuss curvature in Chapter 5, we will recognize that this surface has positive curvature, just from its convexity. If the speed of light were infinite, this surface would be a plane. Likewise, if T is very large (the rotation is very slow), then the curvature is also small. When $T = \infty$, the space is not rotating at all, and the equation of the surface is simply $z = 0$.

From Ptolemy's point of view, each star traverses a circle with period T equal to one sidereal day (approximately 86,164 sec). Considering a near star, traversing a circle of radius 4 light years, which gives $R = 4 \times 366.2422cT$, or $R = 3.76158 \times 10^{16}$ m, one finds the circumference of its orbit contracted to length

$$C = 2\pi r \approx cT\Big(1 - \frac{1}{128\pi^2(366.2422)^2}\Big) \approx 0.99999999409861cT\,.$$

The speed u of the star is less than the speed of light by just $5.90138 \times 10^{-9}\,c$, that is, 1.769 m/sec.

As we shall see in the next chapter, the relativistic mass of this star is about 9200 times its rest mass.

REMARK 1.10. We can express the function z as an elliptic integral of second kind. The substitution $s = (1/2)(3 - \sqrt{3}\tan\theta)$ converts the integral

$$\int_0^u \sqrt{\frac{s^2 - 3s + 3}{(1-s)^3}}\,ds$$

into

$$\frac{3}{\sqrt{2}}\int_{\theta_0}^{\frac{\pi}{3}} \sqrt{\frac{\sec^3\theta}{(\sqrt{3}\sin\theta - \cos\theta)^3}}\,d\theta\,,$$

where $\theta_0 = \arctan\big(3 - 2u)/\sqrt{3}\big)$. For $0 \le u \le 1$ we have $\pi/6 \le \theta_0 \le \pi/3$.

Using the trigonometric identity

$$
\begin{aligned}
\sqrt{3}\sin\theta - \cos\theta &= 2\Big(\frac{\sqrt{3}}{2}\sin\theta - \frac{1}{2}\cos\theta\Big)\\
&= 2\Big(\cos\frac{\pi}{6}\sin\theta - \sin\frac{\pi}{6}\cos\theta\Big)\\
&= 2\sin\Big(\theta - \frac{\pi}{6}\Big),
\end{aligned}
$$

we now change this integral into

$$\frac{3}{4}\int_{\theta_0}^{\frac{\pi}{3}} \frac{1}{\sqrt{\big(\cos\theta\sin(\theta-\pi/6)\big)^3}}\,d\theta\,.$$

Now we can use the trigonometric identity

$$\sin a\cos b = \frac{1}{2}\big(\sin(a+b)+\sin(a-b)\big)$$

to write the integral as

$$\frac{3}{\sqrt{2}}\int_{\theta_0}^{\frac{\pi}{3}} \frac{1}{\sqrt{(\sin\left(2\theta-\pi/6\right)-\sin(\pi/6))^3}}\,d\theta\,.$$

The substitution $\psi = 2\theta - \pi/6$, together with the fact that $\sin\frac{\pi}{6} = 1/2$, then converts the integral to

$$\frac{3}{2\sqrt{2}}\int_{\psi_0}^{\frac{\pi}{2}} \frac{1}{\sqrt{(\sin\psi - 1/2)^3}}\,d\psi = ,$$

where $\psi_0 = 2\theta_0 - \pi/6$.

It is more convenient to integrate in the opposite direction, which we can do by replacing ψ by $\varphi = \pi/2 - \psi$. Letting $\varphi_0 = \pi/2 - \psi_0$, we then have

$$\frac{3}{2\sqrt{2}}\int_{0}^{\varphi_0} \frac{1}{\sqrt{(\cos\varphi - 1/2)^3}}\,d\varphi\,.$$

Next, we use the identity $\cos\varphi = 1 - 2\sin^2(\varphi/2) = 1 - 2\sin^2\eta$ to write this as

$$\frac{3}{\sqrt{2}}\int_{0}^{\eta_0} \frac{1}{\sqrt{(1/2 - 2\sin^2\eta)^3}}\,d\eta\,.$$

Finally, factoring out $2^{-3/2}$ from the denominator, we get

$$6\int_{0}^{\eta_0} \frac{1}{\sqrt{(1-4\sin^2\eta)^3}}\,d\eta\,.$$

If you ask *Mathematica* to evaluate this last integral, you will be told that it is

$$\text{ConditionalExpression}\left[6\left(-\frac{1}{3}\text{EllipticE}\left[\eta_0,4\right] + \frac{2\,\text{Sin}\left[2\eta_0\right]}{3\sqrt{-1+2\,\text{Cos}\left[2\eta_0\right]}}\right),\right.$$
$$\left.\text{Cos}\left[2\eta_0\right] \geq \frac{1}{2}\right].$$

Here, EllipticE $[x,m]$ is the notation *Mathematica* uses for the *elliptic integral of second kind*

$$\int_0^x \sqrt{1 - m\sin^2 t}\, dt\,.$$

The condition imposed on the angle η_0 simply says that $0 \le \eta_0 \le \pi/6$, and this is indeed the case.

This is the first of several times when we shall encounter elliptic functions. They deserve to be better appreciated than they generally are by most physicists and mathematicians. They give exact expressions for solutions to some common differential equations of mathematical physics, solutions that otherwise have to be described qualitatively by inserting corrective terms into expressions involving elementary functions.

14. Problems

PROBLEM 1.1. Solve the equations of the Lorentz transformation for t, x, y, and z in terms of t', x', y', and z', and show that the solution is the same transformation with u replaced by $-u$ (which makes no change in α).

PROBLEM 1.2. Revisit the problem of the twin paradox by imagining that Mary has a telescope trained on the Earth, so that she can constantly observe John's clock. What would she see? Why is it that this clock shows a later time than Mary's own clock when the twins meet at the end of the journey?

PROBLEM 1.3. Consider the vector formulation of the Lorentz transformation given by the mutually inverse relations

$$(t'; \boldsymbol{x}') = \left(\alpha\Big(t - \frac{\boldsymbol{u}\cdot\boldsymbol{x}}{c^2}\Big); \boldsymbol{x} + \Big((\alpha-1)\frac{\boldsymbol{x}\cdot\boldsymbol{u}}{\boldsymbol{u}\cdot\boldsymbol{u}} - \alpha t\Big)\boldsymbol{u}\right)$$

and

$$(t, \boldsymbol{x}) = \left(\alpha\Big(t' + \frac{\boldsymbol{u}\cdot\boldsymbol{x}'}{c^2}\Big); \boldsymbol{x}' + \Big((\alpha-1)\frac{\boldsymbol{x}'\cdot\boldsymbol{u}}{\boldsymbol{u}\cdot\boldsymbol{u}} + \alpha t'\Big)\boldsymbol{u}\right).$$

Verify that these relations really are inverses of each other by inserting the values of $\boldsymbol{x}$ and t from the second relation into the right-hand side of the first relation.

PROBLEM 1.4. Show that the observers O and O' agree about the "space-time metric," that is, show that $(ct^2) - x^2 - y^2 - z^2 = (ct')^2 - x'^2 - y'^2 - z'^2$.

PROBLEM 1.5. Let a and b be dimensionless positive constants. Show that the mutually perpendicular lines $bx = ay$, $z = 0$ and $by = -ax$, $z = 0$ observed by O make the nonobtuse angle $\arccos\big(ab(\alpha^2-1)/\sqrt{(a^2\alpha^2+b^2)(b^2\alpha^2+a^2)}\big)$ when observed by O' at any given instant s. Show that this is a right angle only if $u = 0$, and that it tends to $0°$ as $u \uparrow c$.

PROBLEM 1.6. Translate the equation of the unit circle $x^2+y^2 = R^2$, as seen by O, into O''s coordinate system. What kind of curve does this equation represent? How does the shape depend on time?

PROBLEM 1.7. Translate the equation of a general conic section $Ax^2 + 2Bxy + Cy^2 + Dx + Ey + F = 0$, observed by O, into the coordinate system used by O', getting an equation $A'x'^2 + 2B'x'y' + C'y'^2 + D'x' + E'y' + F' = 0$. Show that the discriminant $\Delta = B^2 - AC$ becomes $\Delta' = (B')^2 - A'C' = \alpha^2(B^2 - AC) = \alpha^2\Delta$. In particular, an ellipse ($\Delta < 0$) remains an ellipse (although, as Problem 1.6 shows, a circle may become a more general ellipse), a parabola ($\Delta = 0$) remains a parabola, and a hyperbola ($\Delta > 0$) remains a hyperbola.

PROBLEM 1.8. Consider the mapping $f(u)$ from $(-c, c)$ onto $(-\infty, \infty)$ given by

$$x = f(u) = \log\left(\frac{c+u}{c-u}\right).$$

(The base of the logarithm is not important here. It may be any positive number except 1.)

Find the inverse of the mapping $f(u)$. Also prove that for the Lorentz addition of velocities $u * v = (u+v)/(1+uv/c^2)$,

$$f(u * v) = f(u) + f(v)\,.$$

Thus, the group of relativistic velocities in one dimension is isomorphic to the additive group of real numbers. (That is no surprise, since, up to isomorphism, this is the only noncompact connected one-dimensional real Lie group that exists.)

PROBLEM 1.9. Suppose that O' observes O and O'' moving away at constant speeds u and v along lines making angle η. Let w be the speed with which O and O'' are moving apart relative to each other, and let

$$\alpha = \frac{c}{\sqrt{c^2 - u^2}}\,; \quad \beta = \frac{c}{\sqrt{c^2 - v^2}}\,; \quad \gamma = \frac{c}{\sqrt{c^2 - w^2}}\,.$$

Show that

$$\gamma = \alpha\beta\Big(1 - \frac{uv\cos\eta}{c^2}\Big)\,.$$

Thus, the angles of a relativistic velocity triangle are determined by its sides u, v, and w via the formula

$$\cos\eta = \frac{c^2}{uv}\left(1 - \frac{\gamma}{\alpha\beta}\right),$$

for the angle opposite side of length w and the analogous formulas for the other two angles.

PROBLEM 1.10. Show that if ξ and η and v and w are interchanged, then Eqs. (1.5), (1.6), and (1.7) remain valid. Likewise, if ζ and η and u and w are interchanged, then Eqs. (1.5), (1.8), and (1.9) remain valid. It follows from Eqs. (1.7) and (1.9) that

$$\frac{\beta v}{\alpha u} = \frac{v\sqrt{1 - u^2/c^2}}{u\sqrt{1 - v^2/c^2}} = \frac{\sin\xi}{\sin\zeta}\,.$$

This relation is the law of sines for a relativistic velocity triangle. Formally, it is the ordinary law of sines applied to a triangle in which the sides opposite the two angles are shrunk, each by the FitzGerald–Lorentz contraction factor that would be measured by the observer corresponding to the opposite vertex.

PROBLEM 1.11. If $\eta = 0$ and $u = v$, then $w = |\boldsymbol{u} - \boldsymbol{v}| = 0$. The expression given by Eq. (1.5) for w shows that it is very unlikely that w^2 can ever equal $|\boldsymbol{u} - \boldsymbol{v}|^2 = u^2 - 2uv\cos\eta + v^2$ in any other case. Because of the denominator $\big(1 - uv\cos\eta/c^2\big)^2$, this certainly cannot happen unless η is an acute angle. Show that this can nevertheless occur at any speed in the case of an "isosceles" triangle corresponding to $u = v$, given a suitable vertex angle. To do so, assume $u = v$ and show that the equation $w = |\boldsymbol{u} - \boldsymbol{v}|$ leads to the quadratic equation $2a^2x^2 - 3x + 1 = 0$ for the unknown $x = \cos\eta$, where a $(0 \leq a < 1)$ is a dimensionless constant, namely $a = u/c = v/c$. The solutions of this equation are

$$x = \frac{3 \pm \sqrt{9 - 8a^2}}{4a^2} = \frac{2}{3 \mp \sqrt{9 - 8a^2}}\,.$$

(You will actually get a cubic equation from which the trivial factor $x - 1$ can be divided out.) Show that the positive sign in the first expression for x (corresponding to the negative sign in the second one) is consistent with the relation $x \in [0, 1]$ only when $a = 1$, which is the case when $u = v = c$. In this case, $x = 1$, that is, this is the case $\eta = 0$, which, as we have already remarked, is trivial. Then, for the negative sign on the square root in the numerator, show that x lies in the range $[1/3, 1/2]$ for all values of $a \in [0, 1]$.

Verify the example mentioned in the text, in which $a = 3\sqrt{2}/5$ and $\cos(\eta) = 5/12$, showing that

$$\eta \approx 67.056553501352011261°$$

and that the relative speed of O and O'' is $0.84c$

PROBLEM 1.12. It is well-known that any three lengths u, v, w with $u \leq v \leq w$ are the sides of a Euclidean triangle provided $u+v > w$. (The philosopher Immanuel Kant cited this fact as an example of what he called *synthetic a priori knowledge.*) Is this true for relativistic velocity triangles? If not, what additional conditions are needed?

PROBLEM 1.13. Show that the sum of the angles of a relativistic velocity triangle is smaller than two right angles.

PROBLEM 1.14. This problem has four parts. We define the *angle defect* of a relativistic velocity triangle whose angles are ξ, η, and ζ to be the positive number $\pi - (\xi + \eta + \zeta)$. Consider a triangle with these angles and divide it into two smaller triangles by drawing a line from the vertex a angle η to a point on the opposite side, thereby dividing the triangle into two smaller triangles, one having angles η_1 (part of angle η), ξ, and φ_1 (at the vertex on the side opposite the angle η), and the other having angles η_2, ζ, and $\varphi_2 = \pi - \varphi_1$.

Part 1: Show that the defect of the original triangle is the sum of the defects of the two triangles into which it is divided. (It is not difficult to prove—although you are not being asked to do so—that when a triangle is partitioned into any number of other triangles, its defect is the sum of the defects of the triangles that partition it.) Thus the defect of a triangle is proportional to what we think of as the area of a triangle, and so we shall define the area of a triangle to be c^2 times its defect. We then define the area of a polygon to be the sum of the areas of any set of triangles into which it can be partitioned. It is not difficult to show that this definition is independent of the way in which the polygon is triangulated.

Part 2: Consider an isosceles relativistic velocity triangle having two equal sides of length u with angle η between them, and let the other two angles both be equal to

ξ and the third side equal to w. Show that

$$\begin{aligned}
\cos\xi &= \frac{\sin\frac{\eta}{2}}{\sqrt{1-\frac{u^2}{c^2}\cos^2\frac{\eta}{2}}}\,,\\
\sin\xi &= \frac{\cos\frac{\eta}{2}}{\alpha\sqrt{1-\frac{u^2}{c^2}\cos^2\frac{\eta}{2}}}\,,\\
w &= \frac{2u\sqrt{1-\frac{u^2}{c^2}\cos\frac{\eta}{2}}}{1-\frac{u^2}{c^2}\cos\eta}\cdot\sin\frac{\eta}{2}\,.
\end{aligned}$$

Part 3: Consider a regular polygon P_n consisting of n isosceles triangles having vertex angle $2\pi/n$ glued together along their equal sides, which all have length u. Show that its perimeter $\pi(P_n)$ and its area $A(p_n)$, which is c^2 times its angle defect, are given by

$$\begin{aligned}
\pi(P_n) &= \frac{2u\sqrt{1-\frac{u^2}{c^2}\cos^2\frac{\pi}{n}}}{1-\frac{u^2}{c^2}\cos\frac{2\pi}{n}}\cdot\Bigl(n\sin\frac{\pi}{n}\Bigr)\,,\\
A(P_n) &= \Biggl((n-2)\pi-2n\arccos\Bigl(\frac{\sin\frac{\pi}{n}}{\sqrt{1-\frac{u^2}{c^2}\cos^2\frac{\pi}{n}}}\Bigr)\Biggr)c^2\,.
\end{aligned}$$

Part 4: Using the relationship

$$\lim_{\eta\to 0}\frac{\sin\eta}{\eta}=1\,,$$

show that

$$\begin{aligned}
\lim_{n\to\infty}\pi(P_n) &= 2\pi\alpha u = 2\pi u+\frac{\pi u^3}{c^2}+\frac{3\pi u^5}{8c^4}+\cdots\,,\\
\lim_{n\to\infty}A(P_n) &= 2\pi c^2(\alpha-1)=\pi\Bigl(u^2+\frac{3u^4}{8c^2}+\cdots\Bigr)\,.
\end{aligned}$$

Thus, for small velocities u the circumference of a circle of radius u is asymptotic to the value it would have it u were a length in Euclidean space, and the same is true of the area. Equality holds if $c=\infty$.

PROBLEM 1.15. Show that the sides u, v, and w of a relativistic velocity triangle given in terms of its angles are all less than c.

PROBLEM 1.16. Compute the relative speed w of O and O'', given that O' has speeds $u=4c/5$ and $v=3c/5$ relative to them and measures the angle between their trajectories as $3\pi/4$.

PROBLEM 1.17. Let ξ, η, and ζ be three positive angles the sum of whose measures is less than π. Show that the expressions

$$\cos^2\xi+\cos^2\eta+\cos^2\zeta+2\cos\xi\cos\eta\cos\zeta-1$$

and

$$\cos\xi+\cos\eta\cos\zeta$$

are both positive, and hence that the formulas given in the text for the sides of a relativistic velocity triangle having these angles are valid.

Problem 1.18. Suppose $(\theta, u, \varphi) \ +_L \ (\chi, v, \psi) = (\mu, w, \nu)$. Prove that for any angles d, e, and f:

$$\begin{aligned}
(d+\theta, u, \varphi) \ +_L \ (\chi, v, \psi) &= (d+\mu, w, \nu)\,,\\
(\theta, u, e+\varphi) \ +_L \ (e+\chi, v, \psi) &= (\mu, w, \nu)\,,\\
(\theta, u, \varphi) \ +_L \ (\chi, v, \psi+f) &= (\mu, w, \nu+f)\,.
\end{aligned}$$

Different operations are being applied to the two addends in each of these cases. Show that if we take $d = e = f$ and combine these results, we obtain a mapping $T_d(\theta, u, \varphi) = (d+\theta, u, \varphi+d)$ that satisfies the equality

$$T_d(\theta, u, \varphi) \ +_L \ T_d(\chi, v, \psi) = T_d\big((\theta, u, \varphi) \ +_L \ (\chi, v, \psi)\big)\,.$$

Thus Lorentz addition is invariant under each operation T_d. (*Caution:* This result does not enable us to reduce the dimension of the three-dimensional space we have invented to describe the addition. We cannot, for example, replace (θ, u, φ) by $(0, u, \varphi-\theta)$ and (χ, v, ψ) by $(0, v, \psi-\chi)$, even though $(\theta, u, \varphi) = T_\theta(0, u, \varphi-\theta)$ and $T_\chi(0, v, \psi-\chi) = (\chi, v, \psi)$. The difficulty is that T_θ is not the same operator as T_χ.

Problem 1.19. Suppressing the second and third spatial dimensions, we focus attention on just the time and first spatial axes. The Lorentz transformation is

$$\begin{aligned}
\tau' &= \alpha\Big(\tau - \frac{u}{c}x\Big),\\
x' &= \alpha\Big(-\frac{u}{c}\tau + x\Big).
\end{aligned}$$

For a reason that will become clear in a moment, let the angle θ be

$$\theta = \arccos\left(\sqrt{1-\frac{u^2}{c^2}}\right) = \arccos\left(\frac{1}{\alpha}\right).$$

Thus, $\alpha = \sec\theta$.

Solve the second equation of the Lorentz equations for x in terms of x' and τ (and θ), then substitute the result in the first equation, so that τ' and x are expressed in terms of τ and x', yielding

$$\begin{aligned}
\tau &= (\cos\theta)\tau' + (\sin\theta)x,\\
x' &= -(\sin\theta)\tau' + (\cos\theta)x.
\end{aligned}$$

Use these equations to solve the car wash puzzle.

Problem 1.20. Here is a variation on the car wash puzzle. Suppose that as the limousine moved through the car wash with speed u, two car wash attendants simultaneously, *as measured by a clock in the car wash*, put scratches in it, one in the front fender, the other in the rear fender. Suppose that they were standing 3 meters apart, *as measured by the car wash attendants themselves*, when they made the scratches. If the limousine is then stopped and measured by the car wash attendants, how far apart will the scratches be?

Problem 1.21. Verify that the space-time interval ds^2 between the two events—the rear of the limousine entering the car wash and the front of it leaving the car wash—is negative (spacelike).

PROBLEM 1.22. Consider the special case of a relativistic velocity triangle when $\boldsymbol{u}$ and $\boldsymbol{v}$ lie along perpendicular directions. For this case, we have $\gamma = \alpha\beta$. Recall, as noted above, that α corresponds to the cosine of the angle of rotation that describes the Lorentz transformation between O and O' when they interchange time coordinates, and likewise β is the cosine of the corresponding angle for the transformation between O' and O'' and γ the angle corresponding to the transformation between O and O''. Show that, when they are regarded as arcs on a sphere, the three angles that provide these geometric representations are the sides of a spherical right triangle in this case.