CLAUSTROPHOBIC PHYSICS: An introduction to the theory of relativity and Poincaré Symmetry

R. Paul Bickerstaff

Physics Department, University of Idaho,
Moscow, ID 83844-0903, USA
and
Physics Department, University of Auckland,
Auckland, New Zealand

Version 0.3: January 12, 1999

These notes are in the process of reconstruction after a disk crash in June 1996, wiped a substantial portion. The reconstruction is being done from printed notes and hand-written corrections. It will take some time to complete. I am deeply grateful to the University of Auckland for providing the facilities that make this reconstruction possible.

(Oct '98: this reconstruction is now basically complete but the final chapter on gravity is undergoing major revision. The rest should be in reasonable shape.)

Dedicated
to
all those
with enquiring minds.

Contents

List of Figures

Preface

The theory of relativity is indelibly associated with the name of Albert Einstein. While many of the basic notions underpinning the theory are remarkably simple and some aspects can be found in the writings of Galileo, and even the Ancient Greeks, it was he who grasped the essence, took it by its horns and shaped the subject into what we know today. Nobody before him had his insight. His great contribution was in revolutionizing our concept of time. Far from being the absolute and universal entity promoted by Newton, he showed that time was inextricably mixed with the location and state of motion of the observer.

Einstein's ideas were so revolutionary that he became a celebrity and grew to epitomize the public's concept of a scientist — as someone dishevelled and a bit eccentric. However, Einstein was anything but your average scientist. He was keenly aware of science's impact on humanity and was by no means an "ivory tower" figure. In fact, he enjoyed a good relationship with the news media and used his public status to promote pacificism and Zionism. In 1952 he was even approached to become President of the young state of Israel. Fame had its price though. His divorce from his first wife included the unusual stipulation that she would get all of the money should he ever win the Nobel Prize, which he did several years later for his work on the photoelectric effect. He also had the misfortune, being of Jewish descent and a public figure, to be a focus of the anti-semitism preached by the Nazis who grabbed power in Germany. His theory was branded as un-German and his books were publicly burned. To their eternal discredit, two German Nobel Prize winners joined in the unscientific criticism. In one album of Nazi enemies his photograph appeared on the first page with the words *Noch Ungehängt* ("not yet hanged") underneath. Unable to return to his home and professorship in Germany from an overseas tour, Einstein settled in the USA where he took up a position at the Institute for Advanced Studies in

Princeton.

Nevertheless, his theory of relativity has successfully weathered criticism from all quarters and now is overwhelmingly supported by experiment. It in fact underpins a very large part of modern science and has reshaped our view of the world. Long regarded as an esoteric branch of knowledge, of relevance only to a few select scientific endeavors, its consequences are rapidly becoming of everyday import as the popularity of GPS receivers catches on.

Einstein's original approach to relativity was through the theory of electromagnetism. It was not long though before people came to realize that a simpler route lay in the consequences of the universal speed of light for mechanics. So, for many years the standard approach has been via study of the classic experiments of Michelson and Morley, which failed to detect any dependence of the speed of light on Earth's motion in its orbit.

In recent times, however, a few have come to realize the logical shortcomings of these approaches, particularly insofar as it is attempted to derive relativity theory from older concepts that are often inconsistent with it. This especially applies to standard treatments of dynamics but criticism can also be levelled at introductory discussions of the constancy of the speed of light. Without synchronized clocks and length standards, speed not only can't be measured, it can't be defined and yet common approaches place discussions of the speed of light before synchronization. Also, modern experiments provide far more convincing evidence in support of relativity theory than do the old experiments of Michelson and Morley. In addition, the picture that emerges in typical introductory courses is a rather limited one, divorced from the modern theoretical physicist's approach using the machinery of Poincaré symmetry.

My attempts at teaching the subject, over several years, have only heightened my own dissatisfaction with the usual presentations. Hence, in this work I attempt an original approach based on a basic postulate of impotence: that an observer in an inertial reference frame can derive no knowledge of his position in space and time, nor of his orientation, nor of his state of uniform motion by conducting experiments within an enclosed laboratory. All of these concepts are relative ones that acquire meaning only when reference is made to one's surroundings. We thus take Poincaré symmetry as a basic postulate and derive everything from it. By this means I have endeavoured to give the subject some of the logical cohesiveness which I believe it deserves. The result has further appeal in avoiding extensive discussions of ether theory (which then has to be demolished immediately it has been taught). Likewise, this approach permits a remarkably simple discussion of the Michelson-Morley experiment — the logical importance of the experiment also emerging in clearer perspective. Perhaps the most noteworthy feature of the present approach though is the treatment of relativistic dynamics. Not only does the connection between symmetry and the conservation laws of energy and momentum emerge without recourse to Lagrangian dynamics but the concept of mass is forced upon us by the Poincaré invariance of isolated processes.

Despite the originality of approach, I believe that the result is an entirely orthodox view of relativity. Part of my discussion on the deflection of starlight might be considered controversial but the presentation is both quantitatively and conceptually correct and my feeling is that the rather weighty criticisms levelled at this type of presentation have been made with too broad a brush. Provided one recognizes the limitations of the simplistic arguments made, there should be no misconceptions generated.

I have used these notes for an introductory (2 credit) course on relativity at the University of Idaho in Spring 1996. This course was listed at the junior level and was intended to be taken after the student had completed a two semester sequence in calculus based physics, including electricity and magnetism. However, courses similar to this are frequently taught at the sophomore level at other universities. For those universities who do not teach such a course, I strongly urge that they consider doing so. (I would recommend a full 3 credit semester course.) Common surveys of relativity within other courses, such as modern physics or electricity and magnetism, can not do justice to the subject and probably do more harm than good. If presented in a manner along the present lines, the subject is not simply another topic in an already crammed curriculum but provides a lean and coherent basis for the higher study of all of physics. It may even be said that relativity and symmetry *is* the basis for physics.

I have tried to keep the presentation as simple as possible without squandering accuracy. An aim has been to write something that would be of wider benefit than a university textbook. The first part on kinematics, which covers the basis of relativity, should be intelligible to students with a modicum of calculus knowledge. It is the concepts rather than the maths which is challenging here and I would like to think that even high school students could follow most of it. Part two, on relativistic dynamics requires a little more mathematical ability but the majority of it should still be intelligible to students with modest mathematical knowledge. As an experiment, I let several students who did not possess the prerequisites enrol in my course.

Their performance was comparable to that of the physics majors and one of these students actually topped the class. All learned enough to have made the course worthwhile taking.

Part of my philosophy in teaching this class has been that one learns relativity by confronting one's misconceptions. I strongly believe therefore that, at the introductory level, ability to resolve problems of a paradoxical nature is far more worthwhile than ability to perform mathematical calculations of time dilations and length contractions etc. My choice of problems and overall presentation reflects this philosophy. Also included are a number of exercises in the text which are meant to further illustrate key points and boost student confidence by doing them. The ends of chapters include revision questions, all of which the student should be able to answer and which instructors should regularly quiz students on. As simple as many of these questions may seem, I have never ceased to be amazed at how little of the basics are absorbed by the average student and instructors should never be reluctant to examine the trivial, for if not mastered it is wasteful to proceed. The end of chapter Questions are meant to stimulate. Not all will be immediately answerable by a student studying this subject for the first time but nevertheless the student should ponder as many of them as possible.

I made these notes available to my class via the world wide web. (The University of Idaho had relatively well equipped student labs in Spring 1996 so web access was not a problem.) Both the content and method of presentation received favorable reaction. I am therefore encouraged to make these notes more widely available, with a view to eventually refining them. I also intend introducing interactive aspects to the web version that will enable readers to explore various issues without getting bogged down in tedious calculations.

Paul Bickerstaff
Moscow, May 1996 Te Atatu South, October 1998

Part I

Kinematics

Chapter 1

Inertial Frames

1.1 Introduction

Have you ever flaked out at a party from overindulgence in liquor, or otherwise been rendered unconscious, and woken up to find yourself in a totally unfamiliar room? You stare at the ceiling and the walls but there is no clue as to where you are. Your watch has stopped and you don't know what the time is. You don't even know which way is North. Only when you look out of the window, or the door, do you learn *from your surroundings* where you have ended up! (If the room had metallic walls, floor and ceiling then it would be magnetically shielded and not even a compass would work. You would still have to go outside to get your bearings.)

Undoubtedly you have at some time been seated in a vehicle and looked out your window at another vehicle passing close by. You get a sudden sensation that you are moving, but the other vehicle passes and you see the buildings and the trees stationary and realize that you are at rest. And when you are moving at a steady speed, is anything inside your vehicle different from when you are at rest? If you drop something, does it not fall straight down as it would if you were at rest?

These simple observations are at the heart of the study on which we are about to embark. If one is in a totally enclosed room, there is no way of knowing where, or when, one is, or whether one is moving uniformly or at rest. In fact, these concepts have no meaning except when used in relationship to one's surroundings.

These ideas are very old. Aristotle discussed time and place at length in his *Physics* [1]. His discussion is very detailed on what it means to talk of the "place," i.e. location, of an object. He notes that place and the object are separate because the object can be moved from one location to another and yet he makes it clear that place has meaning only relative to the surroundings. Thus he notes that a nail in a ship, or water in a vessel, can change its place by virtue of the ship, or vessel, moving, but the nail is still in the ship and the water is still in the vessel. But while he notes that Earth has a place in the heavens, he states that the heavens have no place because there is nothing outside of the heavens (in the dogma of the Ancient Greeks).

The nature of time has also long puzzled man. Aristotle [1] associated it with change and movement. Similar views were espoused by Lucretius who stated [2]:

> Time itself does not exist, but from things themselves there results a sense of what has already taken place, what is now going on, and what is to ensue. It must not be claimed that anyone can sense time by itself apart from the movement of things.

Thus Leibnitz [3] suggested that time should simply be regarded as an order of successions, in contrast to some of his contemporaries, such as Gassendi and Newton, who held that time was absolute and had an independent existence. The view adopted here is that the time depends on what has gone before. When we say that this is the year 1996 AD (or whatever), we are stipulating its relationship to a particular event in the past (in this case, the birth of Christ). If one is not aware of what has happened in the past then "the time" has no meaning.

On the subject of motion, Galileo has given a very explicit account of how things inside the cabin of a moving ship appear just the same as when the ship is at rest [4].

> Shut yourself up with some friend in the main cabin below decks on some large ship, and have with you there some flies, butterflies, and other small flying animals. Have a large bowl of water with some fish in it; hang up a bottle that empties drop by drop into a wide vessel beneath it. With the ship standing still, observe carefully how the little animals fly with equal speed to all sides of the cabin. The fish swim indifferently in all directions; the drops fall into the vessel beneath; and, in throwing something to

> your friend, you need throw it no more strongly in one direction than another, the distances being equal; jumping with your feet together, you pass equal spaces in every direction, When you have observed all of these things carefully (though there is no doubt that when the ship is standing still everything must happen this way), have the ship proceed with any speed you like, so long as the motion is uniform and not fluctuating this way and that. You will discover not the least change in all the effects named, nor could you tell from any of them whether the ship was moving or standing still.

This ignorance of the state of uniform motion is known as the *Principle of Relativity* and was elevated by Einstein to a fundamental postulate of physics. As we have noted though, an observer in an enclosed room is also ignorant of the time and the location and the orientation of the room. In these lectures we shall elevate all of these "ignorances" to fundamental importance.

1.2 Spacetime

Before we discuss these matters further, let us consider for a moment what it is that "where" and "when" are part of. When we specify "where" by giving the location (or place) of an object in reference to other objects, or "when" by giving the time (or epoch) of an *event* (or happening), in reference to other events, we are not saying that the where or when are part of those objects or events. Rather, all of these things may be considered as taking place in an abstract arena. We shall follow modern terminology here by calling this arena *spacetime*. This terminology comes about by appending a time dimension to the familiar three dimensional world in which we live, which is often called space. The three dimensional *space* is just something that is experienced by any observer when they consider moving left or right, forwards or backwards, up or down.

One should be wary of attributing physical properties to either space or time. They are more the mind's creation. One should be particularly wary of divorcing either space or time from the observer. Western thinkers are apt to do this as they are so accustomed to maps. A better way of thinking is that of the ancient Polynesian navigators whose sailing directions were very much self-centered. In order to get from one island to another they would simply note that one should set sail in a certain direction. To get to a third island, one would imagine oneself on the intermediate island and note the direction in which one must head. A map-style bird's eye view of the resultant course was not recognized and, unless the short-cut route was independently known, it would not be sailed. Similarly it is with time. Past, present and future should be associated with the perceptions of an observer. For, when all is said and done, it is only the observer who can verify the truthfulness of our assumptions and hypotheses. This central role of the observer is accentuated by referring to spacetime in terms of the observer's *frame of reference*. This is rather like a mathematician's number line.

We shall therefore assume only the following about spacetime.

Axiom 1 (Spacetime)

1. *Space and time are homogeneous.*
2. *Space is isotropic.*
3. *Spacetime is continuous.*

By the homogeneity of space and time we mean that there is nothing different about either space or time from one place or epoch to another. The asserted isotropy of space adds the further rider that nothing about space distinguishes one direction from another. Continuity of spacetime is meant in the sense that there is no position in space, or instant of time, for which events can't happen. In particular we assert that there is no position in space, or instant in time, at which it is not possible to both emit and receive a flash of light.

Thus, the universe, as we know it, is simply spacetime populated with matter and radiation. Any region of spacetime that is devoid of matter we call a *vacuum*.

1.3 Inertial Observers

We now return to our observer confined in an enclosed room, which we henceforth call his *laboratory*, and add a caveat. We have stated that it is impossible to detect a state of uniform motion from observations conducted entirely within the laboratory. However, what do we mean by uniform? It is certainly possible to detect non-uniform motion. If we are in a ship being tossed by waves on the ocean, then plates will slide across a table as the ship is tipped, and drink in a cup may be spilled. In fact, the motion can make one decidedly seasick. Also, if one is in a plane accelerating down the runway, just prior to takeoff, then one can feel, quite distinctly, that one is being pushed backwards into one's seat.

We eliminate jerkiness and accelerations, which *can* be detected from within an enclosed room, by stipulating that our claims only apply to observers in an *inertial* frame (of reference) which we define as follows.

Definition 1 *An inertial (reference) frame is one in which any object initially at rest with respect to its surroundings, but free of external influence, remains at rest. An* inertial observer *is one in an inertial frame.*

The qualifier, "free of external influence," warrants some comment. In assessing its applicability, a value judgement is required. It is known that any object influences any other through gravitational attraction, even at large distances. But generally the attraction is weak. However, objects charged with electricity, for example, can exert substantial influence on others. In determining whether or not an object is being influenced by its surroundings, or is just reacting to non-uniform motion of the frame, one should look for correlations between the object and its surroundings. If no correlation is found, then the object can be assumed to be free of external influence.

This definition puts in concrete terms our wariness of situations where objects are being tossed back and forth or thrust in a certain direction by virtue of non-uniform movement of the laboratory. It is also a definition readily amenable to practical implementation. Indeed, a device has been built for the TRIAD satellite that corrects for the minute accelerations due to radiation pressure and drag by the residual atmosphere (i.e. wind resistance) at the height of the satellite's orbit. The device basically consists of a small sphere inside a cavity [5]. The position of the sphere is sensed by capacitors and any displacement towards the cavity walls results in a small thrust being applied by a jet on the satellite to recenter the cavity about the sphere. By this means the resultant acceleration experienced by the satellite from other than gravitational effects can be kept to a level of 10^{-11} that of the gravitational acceleration at Earth's surface. To an excellent approximation, this can be considered to be at rest (relative to the remainder of the satellite, which in turn can be considered at rest relative to this free sphere).

One should immediately recognize that an observer on Earth's surface is not inertial. If he releases an object held at rest in his hand then it will not remain at rest, suspended in the air, but will fall to the ground. An inertial observer would be one in a spacecraft in remote outer space, far from any star or planet, where the sensation of weightlessness is being experienced.

An inertial frame can be simulated on Earth by having an observer in free fall. This includes satellites in orbit as well as things simply dropping *freely* to the ground. While in free fall, the observer experiences weightlessness and any object dropped will fall with the same acceleration towards Earth and so appear at rest with respect to the observer.

In practice, a free falling frame is not a perfect inertial one because objects always fall toward the center of the Earth and so objects far apart, moving radially towards Earth's center, will appear to move towards each other. Similarly, an object closer to Earth will experience a greater acceleration (in Earth's frame) than one further away on the same radial line. Such effects can be termed *tidal*. However, on a small scale, tidal effects are minute and can usually be neglected. Under these conditions we refer to the frame as a *locally inertial* one. Strictly speaking, only locally inertial frames exist in our universe.

One last case needs to be considered. We noted that an object released near the surface of the Earth would fall; but suppose something were constraining it. Consider, for example, a puck sliding on a frictionless air table, or a magnetically levitated train gliding along frictionless tracks. Such examples are common and, provided we are only interested in motions of objects which are so constrained (against whatever accelerations would otherwise occur), we are entitled to consider them as *restricted inertial* frames since, because of the constraint, objects at rest remain at rest.

Review

What quantities does an observer in an enclosed room have no knowledge of?

What do we mean by the homogeneity and isotropy of space?

What is the definition of an inertial frame?

Why do we say that some frames are merely locally inertial?

Questions

1. Imagine yourself in a totally enclosed room, satisfying the criteria for an inertial frame. We have asserted that one can not determine the location, epoch, orientation or state of uniform motion of that room. Would you be able to determine any difference

   (a) between your room and a mirror image of that room?

   (b) if the direction of time were reversed, i.e. if all motions were reversed and everything were to run backwards?

   (c) if the scale of the room (and everything in it) were to change, either shrink or expand?

2. One can move in any of the three spatial dimensions in either a forwards or reverse direction. Thus one can move right or left, and one can move towards the front or the rear, and one can move up or down. However, it seems that one can only move forwards in time and not backwards. Why is that?

3. A vacuum has been defined as a region of spacetime devoid of matter. What is matter? How would you detect its presence or absence? (Do we really need to know what matter is in order to establish a vacuum or merely how it behaves?)

Problems

1. A skydiver jumps out of an airplane. Is the skydiver in an inertial frame, prior to his parachute opening? Explain clearly why or why not.

2. We have noted that an object dropped near Earth's surface will fall. However, a helium filled balloon will rise because helium is lighter than air and the net density of the balloon and helium causes it to float upwards. Suppose we fill the balloon with a mix of air and helium so that the net density of balloon and gases inside it is exactly the same as air. Then, this balloon when released will float motionless. Does this mean that the frame is inertial (or restricted inertial)?

Project: The Ancient Greeks

Locate a copy of the works of Aristotle. (One place to try is the Encyclopedia Britannica series, *Great Books of Western Civilization*.) Browse through his writings and read some of them in depth.

1. Find out a little about Aristotle. When did he live? What events were taking place in his time? What was his relationship to other Greek philosophers?

2. Aristotle was held in extremely high regard in the Middle Ages as an impeccable authority but in recent centuries his stature has been greatly diminished. Many modern physics texts have vilified him for some errors he made, such as on the flight of arrows. Based on your reading of his work, do you consider him more a fool or a very wise man? Should a person's ideas be wholely dismissed because of a few mistakes? (You may care to ponder how a reputation is harder to build than destroy.)

3. Aristotle often refers to the ideas of earlier philosophers, sometimes agreeing with them and sometimes criticizing harshly. Find some examples of this. How important is it for any intellectual endeavor to build on the knowledge of others?

4. Read the sections in Book IV of Aristotle's *Physics* on time and place. Discuss how closely his ideas gel with modern notions of these issues. Was he right in every respect? When he was right, do you think he merely fluked it or did he arrive at those conclusions through sound reasoning? Even if his reasoning was sound, is reasoning alone enough? (Consider that great minds such as Newton have arrived at different conclusions.)

5. What role did experiment play in Aristotle's philosophy? (Be careful to distinguish between keen observation and purposeful experiment.) Without modern experiments to guide us, would Aristotle's views on time be any more worthy than the contrary views of Newton and Gassendi?

Project: Testing for Inertial Frames

1. Provide draft plans, or a brief but concise description, for at least 2 different devices for detecting whether or not one is in an inertial reference frame.

2. Judge these devices based on criteria of accuracy, cost, ease of construction and any other criteria you deem relevant.

3. Based on one or more relevant criteria, select one of your devices and build a working example.

Chapter 2

Measurement of Time and Space

2.1 Time

Measurement of time is accomplished with an instrument called a *clock*. A clock can be anything that exhibits a uniform periodicity, or changes in a uniform way.

Perhaps the oldest clock is the spinning Earth. As Earth rotates, the side facing the Sun changes, leading to the apparent movement of the Sun across the sky and the onset of night. This leads to the interval of time known as a day. Similarly, the changing seasons as Earth moves around the Sun in its orbit gives rise to the year. For some primitive peoples it was quite satisfactory to divide up the day into such things as early morning, mid-day, late afternoon and night, but as societies became more rigid in their behavior, the need developed to divide up the day more precisely. The modern convention is to divide it into 24 hours and an hour into 60 minutes and the minute into 60 seconds. The division into 60 stems from the ancient Babylonians, who attributed mystical significance to this number. Why the day was divided into 24 hours has been lost in antiquity. One possibility is that ancient peoples measured the height of the Sun in the sky by holding their hand outstretched in front of their faces and marking off the number of spans. It is known that traditional Arab navigators used this technique to measure the altitude of stars above the horizon. It just so happens that an outstretched hand subtends an angle of close to 15° at the eye and with 360°

in a full circle we have that $360/15 = 24$ such spans would make a full day.

The time as measured by the Sun can be made more precise using an instrument known as a sundial. This assumes that the sun moves across the sky at a fixed rate and the shadow cast by an object is compared with a graduated scale. However, the rate of movement of the Sun turns out to vary with the seasons. This variability of "solar time" is due to the same $23°$ inclination of Earth's spin axis to a perpendicular to the plane of Earth's orbit around the Sun that causes the seasons. One easily sees that if Earth's spin axis were in the plane of its orbit then there would be no daily periodic movement of the Sun in the sky at all, whereas if the spin axis were perpendicular to the plane of the orbit then the rate of daily movement would not change during the year. The variation in rate during the year is due to the intermediate inclination. A more accurate method of measuring time is to make reference to the movement of the distant stars. This leads to sidereal time. For very accurate measurement though, even this has its drawbacks because it is known that there are variations, due to tidal effects, in the rate at which Earth spins [6]. Tidal friction is also slowing down Earth's rotation rate. Though the lengthening of an individual day is only about 2s every 100,000 years, the accumulated effect grows quadratically and is of the order of half a minute in a century and several hours in a few millennia [6]. This is readily observable in astronomical observations and indeed is highly significant in analyzing eclipse observations from antiquity [7].

There are many other types of clock such as those employing the oscillation of a pendulum. Many ingenious methods were once devised to make pendulum clocks more accurate. One can also construct satisfactory clocks by using things which are not periodic at all but which change at a constant rate. Examples are the hour glass (in which grains of sand fall through a narrow glass neck), or the Roman water clock (consisting of a vessel of water with a hole in it through which the water slowly drained out).

One particular constant-rate phenomenon which can be used as a clock is radioactive decay. It is found that certain atomic nuclei and subatomic particles decay in such a way that the rate of decays is proportional to the amount of undecayed material. This can be expressed mathematically as an exponential decay law and has the feature that in equal time intervals the number of undecayed particles decreases by a constant factor. When this factor is $\frac{1}{2}$ we call the time interval the half-life of the decaying particles. Time can thus be measured in terms of the number of half-lives and so this system constitutes a valid clock.

For accurate measurements we must have a very reliable clock that can be readily duplicated for use in laboratories the world over. The most accurate clock available at present is the so-called atomic clock. This is based on the fact that in transitions between different atomic states, radiation of a well-defined periodicity is emitted and the period is a characteristic of that transition in that atom. Similarly, atoms absorb radiation at the same well-defined periodicity. A typical way of utilizing this to make a clock is to pass electromagnetic radiation, of about the right periodicity, through a gas of particular atoms and fine-tune the periodicity by a feedback mechanism so that absorption occurs.

Such a highly accurate clock can be dubbed a standard clock, against which all others must be calibrated and corrected. Since 1967, the internationally adopted standard clock is an atomic clock based on the radiation emitted during a transition between the so-called hyperfine levels of the fundamental state of the atom cesium 133 (^{133}Cs). The cesium-beam atomic clock is accurate to about 2 parts in 10^{14}. This amounts to less than one second of error in a million years. Accordingly, the second is now *defined* as the duration of 9,192,631,770 periods of the above mentioned radiation. This number was chosen to correspond as closely as possible to the previous standard.

2.2 Straight lines

Before we can measure distances we need to resolve a small problem. It is clear that the distance between two points depends on the path travelled between them. We would like to define the length of an object as the straight line distance between its endpoints. But what precisely is a straight line?

Euclid defined a straight line, in his classic treatise [8] on geometry, as "a line which lies evenly with the points on itself." If that seems rather obscure to you, don't be disheartened! There has been much debate in scholarly circles over the centuries as to exactly what he meant. The view of modern geometers is that in geometry there must always be some primitive notions that can't be well defined and these only acquire full meaning after postulates and theorems involving them have elucidated their character. However, this is not of much use to a physicist who desires to know whether he is measuring the length of an object or something else. Plato [9] remarks that a straight line is one in which the points are hidden by the ends, thus suggesting that

Euclid may have meant that a straight line was to be judged by peering along it and seeing if any parts jutted out. This is in fact a common way of judging whether or not something is straight.

We therefore shall adopt the following definition.

Definition 2 *A straight line in space is that path in an inertial frame that is travelled by light in a vacuum.*

The qualification that light travels in a vacuum is to avoid the complication of refraction. The frame clearly needs to be an inertial one because a straight line in an inertial frame is not going to be straight in a non-inertial one.

2.3 Light

Since light has seized an essential role in our discussion it behoves us to consider briefly something about its nature. We know that light possesses wave-like properties because it exhibits interference and diffraction phenomena. However, it also can exhibit particle-like properties such as in the photoelectric effect, or Compton scattering, where energy is exchanged in quanta which we call photons. We don't really understand what light is though and it won't be necessary here to try and do so.

Of more direct relevance are some of the essential features of light. We know that light can propagate through a vacuum and that it travels with a finite speed, even if that is very fast.

The first estimate of the speed of light was made by the Danish astronomer Roemer, in 1675. Roemer had been observing Jupiter and its moons and noticed that the times of occultations of the moons, by the planet, differed by up to about 16 minutes during the course of a year. The motion of the moons in their orbits could be assumed to be highly regular so Roemer deduced that the variation in observed times of occultation must be due to the light from the moons having to travel farther. In the course of 6 months, Jupiter only completes a small fraction of its orbit whereas Earth completes one half. Thus the maximum variation in times occurs due to the light having to travel the diameter of Earth's orbit about the Sun. See fig. 2.1. Approximate knowledge of the diameter of Earth's orbit allowed Roemer to determine the speed of light.

Other early measurements included Earth-bound experiments using such devices as rapidly rotating mirrors on hills separated by several miles. One

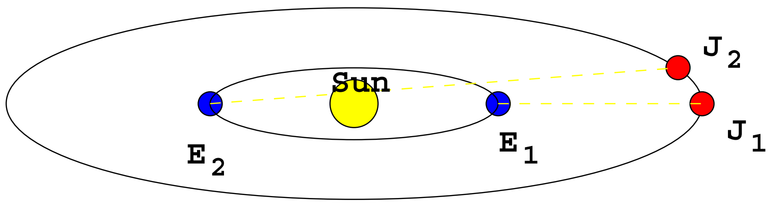

Figure 2.1: As Jupiter and Earth move from positions 1 to 2, the distance that light from Jupiter must travel to reach Earth varies by the diameter of Earth's orbit.

measurement of this type was conducted by Foucault in 1846. When the British physicist James Clerk Maxwell discovered that electromagnetic waves were predicted by his theory of electricity and magnetism, their velocity (given by electromagnetic constants) turned out to be so close to the then known value of the speed of light that he immediately guessed that light was this type of wave.

Nowadays, the finite propagation time of light can be measured quite accurately. For example, astronauts have left a corner reflector on the Moon. This simple device has the property that light is always reflected back in the direction from whence it came and has been used to bounce laser pulses off of the Moon. The round trip travel time is found to be of the order of 2.5 seconds.

Maxwell's theory of electromagnetic waves asserts that light travels at the same speed in vacuum regardless of wavelength. Thus, radio waves, microwaves and gamma rays which are all forms of light, merely lying outside of the visible wavelength range, are expected to travel at the same speed. As far as we can tell, they do. If they did not then we would see different images of the planets and stars, one image for each color. However, we don't; all colors come from the same spot in the sky, as do any radio emissions. (The rate at which light travels inside materials certainly does depend on wavelength and this is responsible for the phenomenon of refraction.)

2.4 Distance

There are several ways we can envisage measuring distance. Perhaps the simplest is to take a measuring rod, such as a foot-rule, or yardstick, and measure off the distance by placing it end on end and counting the number of multiples of its length there are in the distance being measured. In carrying out this procedure, we assume that the length of the measuring rod doesn't change as we move it from one place to another. We also assume the length is independent of the orientation of the rod and is the same today as it was yesterday.

There are though a number of drawbacks with this method. One is that its practical application is limited; one can't measure the distance to the Moon by laying out yardsticks. This is not a severe problem however as we can use trigonometry to determine the distance from other known distances and angles subtended.

A more serious problem concerns accuracy. Consider for example the case of the meter. Originally defined as one ten millionth of the distance along Earth's surface between the equator and the North pole, measured through a line of latitude passing through Paris, imprecise knowledge of that distance led to a standard meter being marked on a bar of platinum kept in Paris. However, this was both inconvenient and inaccurate, as the bar would expand and shrink slightly with temperature changes and had to be kept under specially controlled conditions. This only made it even less accessible.

There are other common ways to measure distance. For example, we might say that it is a half-hour walk to school, or a five hour drive to Seattle. If the speed of walking, or driving, can be standardized then an accurate measure of distance by this means is possible. Of course everybody walks or drives at somewhat different speeds and so we must agree on something that we believe should move at the same speed for all inertial observers. For want of something better we choose light.

Hence we can define distance by using a standard clock to time light during a round trip. By placing a mirror at the destination point we can sit at the starting point with just a single clock. The method has the advantage that we automatically are determining the straight-line distance. Thus we can define the length of an object as follows.

Definition 3 (length) *The length, l, of an object is the (straight line) distance between its endpoints, determined by measuring the round trip time $\Delta\tau$ for light to travel between the ends. Thus*

$$l = c\Delta\tau/2 \tag{2.1}$$

where c is an arbitrarily chosen number, known as the speed of light, which serves to define length units in terms of time units.

It should be clear that the object being measured must be at rest with respect to the observer and his light source (and mirror), for if it were moving then the far end would have moved before the light pulse got there and an inaccuracy would be introduced.

The high accuracy of atomic clocks enables more accurate length measurements than can be achieved with measuring rods and trigonometry. For example, the Earth to Moon distance is now known to an accuracy of a few centimeters and radar ranging with spacecraft visiting other planets has led to a significant revision of the distance to the Sun and planets.

As an example of the units, suppose we chose $c = 1$ light-year/year. This determines the unit of length called a "light-year" (ly) as the distance travelled by light in one year. Similarly we could define the "light-sec" by choosing $c = 1$ light-sec/second. If we chose c to be just the number 1, without any units, then we could measure both time and distance using seconds.

In 1983, it was decided at an official Conference on Weights and Measures to redefine the meter as the distance travelled by light in vacuum in $1/299,792,458$ of a second, i.e. the Conference *defined*

$$c = 299,792,458 \,\text{meters/sec}. \tag{2.2}$$

For many purposes it will be adequate to approximate this by $c = 3.00 \times 10^8$ m/s.

Because c is defined to have a precise value, it makes no sense to try and measure it more accurately. One could only measure the speed of light if one had a length standard that was independent of it, such as the old platinum bar in Paris. An immediate question arises here as to how we know that the speed of light is the same always, because the definition implicitly assumes that it is. Well, we don't. What we have done is transferred all uncertainty in this matter to the length measurement. The possibility thus exists that an object might have a different length in a different place, or if it is rotated, etc. However, this assumption was always there anyway with the measuring rod method. Things are no different now and certainly no worse. We shall return to this point in the next chapter.

One will note that, because of the high speed of light, distances on the scale of the human body are covered in a very short time. It helps to develop a feel for the scales involved. An interesting fact is that in 1 nanosecond (1 ns $\equiv 10^{-9}$ s) light travels quite close to 1 foot.

Exercise 2.1 *Determine exactly how far, in inches, light travels in this time. Note that 1 inch equals 2.54 cm. (This conversion factor has been decreed by acts of both the British Parliament and the US Congress!) There are, of course, 12 inches in 1 foot.*

2.5 Synchronization

We are now in a position to define the (average) speed of moving objects as

$$v = \Delta l / \Delta t$$

where Δl is the distance travelled and Δt is the time taken. However, different procedures for making these measurements could be envisaged. (For the moment let us assume travel along a straight line in such a manner that the traveller perceives the inside of the vehicle to be an inertial frame.)

If Δt is measured by an observer travelling with the object, via his wristwatch for example, then the starting and finishing points are moving in his frame and the distance being travelled is not accurately measured by the method implicit in the definition of length. On the other hand, if the time is being measured by an observer at rest with respect to the starting and finishing points then more than one clock is required. One needs a clock at the starting point to record the time of departure, t_1, and a clock at the finishing point to record the time of arrival, t_2. But $\Delta t = t_2 - t_1$ is not the time of travel unless the two clocks are synchronized.

The issue of synchronizing clocks turns out to be rather subtle when very accurate measurements are required. A popular notion of synchronization is to compare two clocks at the same place and transport them after adjusting them to be equal. Unfortunately, modern experiments have shown this method to be invalid. If a clock is moved, it runs at a different rate than when it is at rest. It also runs at a different rate at the top of a mountain than at the foot.

A convincing experiment showing this was performed in 1971 by Hafele and Keating [10]. They took a pair of very accurate atomic clocks, of identical construction, and flew one around the world while the other remained on the ground. When the clocks were compared at the end of the journey they were found to differ by several hundred nanoseconds. A similar experiment was performed a few years later, in which an atomic clock was flown in circles above Chesapeake Bay for several hours. The same result was obtained: the two clocks kept different times.

There is also very strong evidence from particle accelerator laboratories and cosmic rays that the decay of unstable subatomic particles appears to take place at a slower rate than when they are at rest with respect to the observer. Hence we can conclude that even standard clocks can not be assumed to keep the same rate unless they are at rest with respect to each other and are in an inertial frame.

In order to synchronize two clocks we shall follow a procedure suggested by Einstein. Send a light signal from one clock to the other and have it reflected back. Record the time of departure, $t_1^{departure}$, on the first clock, the time, $t_2^{halfway}$, on the second clock that the light reaches it and the time, t_1^{return}, on the first that the light returns. If

$$t_2^{halfway} - t_1^{departure} = t_1^{return} - t_2^{halfway} \tag{2.3}$$

then the clocks are said to be synchronized. Simple algebraic manipulation of this equation shows that

$$t_2^{halfway} = (t_1^{return} + t_1^{departure})/2. \tag{2.4}$$

This result is known as the *radar rule*. We are in effect making an assumption that, according to clock 1, the light arrives at clock 2 exactly halfway between the time of its departure and return. This of course is something that clock 1 really doesn't know about but we nevertheless assign this time to the event. If clock 2 reads this same time when the light reaches it then the clocks are synchronized. Basically, we are making an assumption here that the light is travelling at the same rate in both directions. However, this is not verifiable; even if we established an independent length standard, the speed of light in each direction could not be measured without using synchronized clocks.

2.6 Proper time and length

Having synchronized our clocks, it is now possible for an observer at rest with respect to the starting and finishing points to measure the time of travel and determine the speed. We should however distinguish this time measurement from that made by the traveller since the latter has the distinction of being performed with just a single clock. Such time we call the *proper time*, τ.

If more than one clock is required then we call it just the time, t, or the *coordinate time* — the latter terminology following from the discussion of the next section.

In order to have a consistent definition of speed, the distance measurement should be made in the same (inertial) frame as the time measurement. To measure the length of a moving object we stipulate that its endpoints be marked off simultaneously on a stationary rod and the distance between the marks be measured by the standard means. When the object whose length is to be measured is at rest with respect to the observer, we call the length the *proper length*, l_0. Otherwise it is just the ordinary length, l.

Thus a person travelling in a car between two cities will measure the proper time, $\Delta\tau$, (using her wristwatch) but the distance measured between the cities will be the ordinary length, Δl, of the moving road. Clearly the length measurement poses some practical difficulties for the traveller but in principle it could be done. (Think of a very long train instead of a car. Let the train be so long that it reaches from one city to the other. Have a set of observers in each carriage, equipped with synchronized clocks, and have them record the time they pass each railway station. Then have them communicate their results to a central processing facility. At this facility, select a pair of recordings, one from each end of the train, that were made simultaneously. Then, the proper distance between the two observers, at rest with respect to each other and the train, is the ordinary distance between the railway stations.) Thus, in the traveller's (restricted inertial) frame the speed at which the road passes by is

$$v_{road} = \Delta l / \Delta\tau \tag{2.5}$$

On the other hand, an observer at rest with respect to the road will measure the proper distance, Δl_0, between the cities but needs two clocks, one to record the time of departure and one to record the time of arrival. Thus that observer measures the ordinary time, Δt, and calculates the speed of the traveller's car as

$$v_{car} = \Delta l_0 / \Delta t \tag{2.6}$$

2.7 Coordinates

In order to make lots of time and distance measurements between different points in spacetime it helps to have a system of coordinates, much like one might use a graph-paper-like grid to locate points in space. It must be stressed though that this coordinate system will be our construction and is not something that is inherent in spacetime. No coordinate grid exists on Earth's surface and if we wish to set up lines of latitude and longitude, or if we just wish to establish a grid on paper for drawing graphs, we must devise a means to construct it. Such considerations led to the science of geometry, which literally means land measurement ("geo" is Greek for land and "meter" is Greek for measure). According to tradition, geometry had its roots in Ancient Egypt. After the annual floods of the Nile, it was necessary to resurvey all of the plots of land in order to accord correct ownership. Just as the plots of land in Ancient Egypt were not a natural feature of the land, but were assigned, so our coordinate system for spacetime will be assigned.

One further point should be emphasized. We are going to attempt to represent both time and space dimensions on a sheet of paper. It should be clear that this is *not* the same as actually constructing a coordinate system in spacetime.

Our basic procedure is to place a set of synchronized clocks at regular intervals in space. An *origin* is arbitrarily chosen, at which the time and space coordinates are selected to be zero, and the space and time coordinates of any event are determined by reference to the location of the nearest clocks and the time they read. In principle, we can erect a lattice of measuring rods in space and at each calibrated mark on the rods we place a clock. In practice we don't need to actually do this; we just use that portion of the lattice necessary to measure the desired coordinates and this should only involve a handful of rods and clocks. Nevertheless, the entire lattice can be imagined to be in place.

Let us now consider the representation of this on paper. Since we can't make marks on the paper change with time we must represent different times by different places on the paper. This makes the representation of all four spacetime dimensions essentially impossible. However, we can easily represent one space and one time dimension. Let time progress towards the top of the page and the space direction increase to the right, as in fig. 2.2.

Our procedure for synchronizing clocks mean that, as represented on paper, the light signal involved in the procedure traces out two sides of an isosceles triangle. Furthermore, the line of simultaneity must make a right angle with the time axis, because the triangle is isosceles and because of the equality of the time intervals in Eq. 2.3.

Because all such lines of simultaneity make right angles with the time axis,

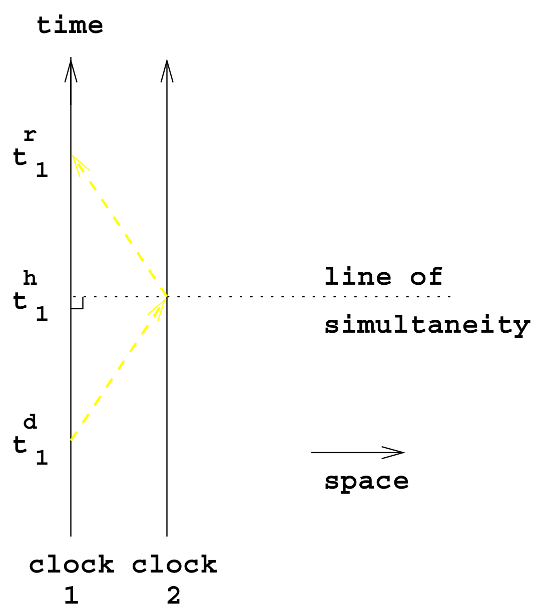

Figure 2.2: The radar rule for synchronizing clocks. A light pulse is emitted at time t_1^d according to clock 1 and returns at time t_1^r according to clock 1, after having been reflected at the position of clock 2. The time on clock 1 at the instant the light pulse is reflected at 2 is decreed to be halfway between the time of departure and return.

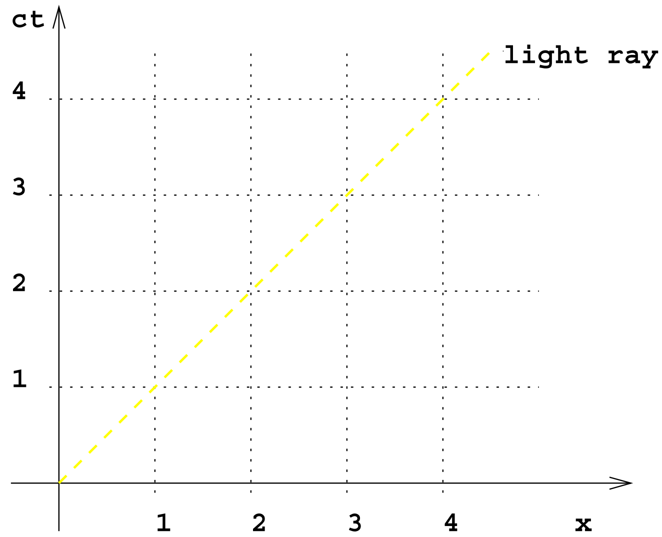

Figure 2.3: Time-space coordinate grid. The distance travelled by a light pulse (dashed line) is proportional to the time it takes. For c, the speed of light, the pulse traces a line of slope 1 on a ct versus x, y, or z plot.

they will be parallel to each other. The homogeneity of time furthermore implies that equally spaced time intervals should be equally spaced on the paper.

Now recall the definition of distance. Given that c is fixed and time is homogeneous, distances do not change with time and so a line of constant distance from the origin must be parallel (in the time dimension) to the time axis (i.e. the line traced out by a clock at the origin). Because of the homogeneity of space, all such lines representing the time evolution of equally spaced marks on a measuring rod, should be equally spaced on our paper.

If we arbitrarily choose units in which $c = 1$, or use instead of t the combination ct, then we get a square coordinate grid as in fig. 2.3. Note that on this grid, a light ray traces out a line of constant slope of 1 (i.e. it is at 45° to the axes).

Such a coordinate system can be constructed for any inertial observer. There is no reason to suppose though that the construction will lead to the same grid. The coordinate system, by construction, is something that belongs to a particular observer, and his inertial frame. It does not belong to spacetime.

Having constructed the coordinate system we specify the location of any event in spacetime in terms of its time and space coordinates. We write these as a 4-tuple: (ct, x, y, z) where x, y and z are the usual space coordinates and t is the time. This 4-tuple may also be written as (x^0, x^1, x^2, x^3) and may be regarded as a four-component vector in the four dimensional spacetime.

Exercise 2.2 *Sketch a time-space coordinate system for an observer and mark on it the events involved in measuring the length of a rod, one end of which is located at the origin. (Take the rod to be oriented along the x-axis.)*

Review

What is a clock?

What is a straight line?

How long does it take light to make a round trip from Earth to Moon; or a one way trip from Sun to Earth?

How does one synchronize two clocks?

What is the radar rule?

What is proper time?

What is proper length?

Is a coordinate system a property of spacetime or an observer dependent construction?

Questions

1. It is often said that the shortest distance between two points is a straight line. However, it is well known that the shortest distance between two places on Earth's surface is a so-called "great circle." Thus a plane travelling from the USA to Europe (which are roughly at the same latitude) will not fly directly across the Atlantic ocean but will first fly north and pass close to the North Pole before heading south. Is this "great circle" a straight line? Why?

2. We have noted that an accelerated clock does not keep time at the same rate as one at rest in an inertial frame. It is not possible therefore to (accurately) synchronize two clocks by bringing one to the location of the other, adjusting it to be the same, and then taking it back. However, could we synchronize two clocks by bringing both to a midpoint, setting their time to be the same and them returning both to their original positions with identical accelerations? (How would we determine the midpoint? How would we guarantee that the accelerations during return were identical?) Could one use this method to synchronize a whole lattice of clocks?

3. Consider Roemer's determination of the speed of light using the occultation times of the moons of Jupiter. Jupiter takes approximately 12 Earth years to complete one orbit. Suppose we were to repeat Roemer's measurement 6 years apart. Then light from Jupiter would be reaching Earth from opposite directions. Would Roemer's procedure therefore provide a means to measure the one-way speed of light?

4. The construction of a coordinate system implicitly assumes that a clock can be located at a point. How small can a clock be?

5. Is there any reason why v_{road} in Eq. 2.5 and v_{car} in Eq. 2.6 should be related?

Problems

1. In this problem we analyze the slowing rate of rotation of planet Earth.

   (a) The rotation rate (or angular velocity), ω, of a rotating object is defined as the angle $\Delta\theta$ through which it turns in a time interval Δt, i.e.

   $$\omega = \frac{\Delta\theta}{\Delta t}.$$

   Show that the rotation rate of the Earth is 15°/hour, or 15 arc-sec/s.

(b) Tidal friction is slowing down the rate of Earth's rotation at a constant rate

$$\alpha = \frac{\Delta\omega}{\Delta t} = -0.99 \times 10^{-16} \text{ arcsec/s}^2.$$

i. Show that the Earth's rate of rotation 100,000 years ago was approximately 30 arcsec/day faster than now and confirm that this implies that the day then was about 2s shorter than now, as stated in the text.

ii. How long, according to modern time standards, was a day 150 million years ago, during the Jurassic period, when dinosaurs such as apatosaurus, brachiosaurus and diplodocus roamed the Earth?

(c) Show that the total angle turned by the slowing Earth in time t is

$$\theta = \omega_0 t + \frac{1}{2}\alpha t^2$$

where ω_0 is the initial rate of rotation. Conclude therefore that the difference in angles turned by a slowing down Earth and a constantly turning Earth is

$$\Delta\theta = \theta - \omega_0 t = \frac{1}{2}\alpha t^2.$$

Hence show that in one century, $\Delta\theta \approx 550$ arcsec and that this corresponds to a difference of about half a minute in the times of sunrise and sunset.

(d) Plutarch notes an eclipse "which, beginning just after noon, showing us plainly many stars in all parts of the heavens, and produced a chill in the temperature like that of twilight." This eclipse took place on March 20, AD 71 and we may conclude from Plutarch's description that it was close to total. A total eclipse never lasts more than 8 minutes. By how much would the time of noon on that day differ from that if the Earth had been spinning at the same constant rate as in 1996?

(e) An eclipse of the Sun is caused by the shadow of the Moon as it passes between Sun and Earth. The width of this shadow on Earth is never more than 269 km across and the path of the shadow defines the places where a total eclipse of the Sun is observed. Herodotus records that an eclipse took place during a battle between the Lydians and the Medes. He states that "the sun assumed the shape of a crescent and became full again, and during the eclipse some stars became visible." This spooked the superstitious combatants so much that they made peace, thus concluding a five-year war. It is possible to conclude, by backtracking the motions of Earth and Moon, that this eclipse took place on August 3, 431 BC and was total in parts of Asia Minor. Deduce however, that if Earth were not slowing down then the "path of totality" would have passed so far away that an eclipse so close to totality would not have been observed at that particular place. (The radius of the Earth is about 6.37×10^3 km.)

2. Determine the number of meters in a light-year. Begin by taking a year to be precisely 365 days. A tropical year (defined from equinox to equinox) is 31,556,925.2 s long. How many days is this? What is the difference in the length of this light year and the naive one calculated earlier? One can also use a sidereal year (defined in terms of the fixed stars) which is 31,558,149.8 s long. How long a light-year would this give? Can you suggest why professional astronomers prefer not to use the light-year as a distance unit?

3. Show that the (proper) length of a stationary object, defined in Eq. 2.1 using a round trip light signal, is the same as the length, $l = c\Delta t$, determined using a pair of synchronized clocks and a one-way light signal.

4. A commonly advocated way of synchronizing a pair of clocks is to measure the distance L between them and to prearrange that a light signal will be sent to clock 2 from clock 1 at time t_1 according to clock 1. Clock 2 is then to be set at time

$$t_2 = t_1 + \frac{L}{c}$$

at the instant the light signal arrives.

   (a) Prove that this is a valid method.

(b) Discuss carefully the measurement of L. Does this method of synchronization use a one-way speed of light?

5. It is implicit in our discussion of synchronization that if two standard clocks are synchronized then they must remain synchronized. Why?

   (a) Show that if two light signals are sent from clock 1 at times t_1 and t'_1 (according to clock 1) and are received at a second clock, synchronized with the first, at times t_2 and t'_2 respectively (according to the second clock) then

   $$t_2 - t_1 = t'_2 - t'_1.$$

   (b) Is the converse true? (That is, if the times recorded on the two clocks for light signals travelling from clock 1 to clock 2 always satisfy the above equation, are the two clocks synchronized? Be warned: the wrong answer has been published in a refereed physics journal.)

6. Prove that if two ("slave") clocks are each synchronized with a third ("master") clock then they are synchronized with each other. (Be sure to treat the general case when the clocks are not colinear. Does the general case require Euclidean geometry?)

Project: Accurate clocks

Select one type of clock.

1. Find out what you can of the history of this clock and the use made of it. Does it have any limitations on when or where it can be used? Is it still in use? (Record references for all sources from which you gather information.)

2. Build at least two versions of this clock and conduct a series of runs in which the times they tell are compared with each other. Discuss the observed variance in terms of the need for standard clocks to not only be reliable but also replicable.

3. Analyze this clock. Explain both the basic mechanisms that make it suitable as a clock (i.e some periodic or steady rate phenomenon) and the factors which limit its accuracy.

4. Suggest some way in which the accuracy of your clocks might be improved (and their variability reduced). Attempt, if possible, to make this improvement and verify that your modifications are indeed successful.

5. The accurate determination of longitude is of critical importance to navigators. One approach is to use celestial observations in conjunction with time.

   (a) Show that the time at which a star passes overhead can be used to pinpoint one's longitude (if a common time can be established over the Earth's surface). Show though that an error of 1 minute in the time leads to an error in position of about 28km if one is at the equator. (This is quite enough to lead to dangerous uncertainty with regard to the location of hazards such as reefs.) [Earth's radius, R, is 6.37×10^3 km and it rotates once in 24 hr.]

   (b) In 1714 the British government offered a prize of 20,000 pounds (a huge sum in those days) to the first person to construct a clock capable of measuring longitude to within half a degree at the end of a voyage from England to Jamaica. [In those days the voyage took 6 weeks and the required accuracy in longitude amounts to 48 km in position at the latitude of Jamaica.] The prize was won by John Harrison for his "chronometer". In 1761 one of his chronometers was found to be only 5.1 seconds slow after 81 days of rough sailing. A chronometer carried by Captain Cook on his second voyage (1772-1775) lost just 8 minutes in 3 years.

       How does your selected clock compare with Harrison's chronometer? Might you have been 20,000 pounds richer if you'd lived in the 18th century?

Chapter 3

The Invariance of Physics and Newton's First Law

3.1 The Invariance of Physics

Now that we have decreed procedures for measuring space and time, and have defined all that is necessary to talk meaningfully about such things as uniform motion in a straight line, we can proceed to formalize the notions discussed in chapter 1.

In chapter 1 we observed that a person in an enclosed room knew nothing of the outside and that concepts such as the location and time of an event, or the orientation of an object, or its state of uniform motion in a straight line, only had meaning in terms of the surroundings. We now raise this observation to the status of a postulate; one on which all of physics will be founded.

Postulate 1 (Invariance of Physics) *It is not possible to conduct any experiment within an enclosed laboratory, in any inertial frame, that can distinguish the location, epoch, orientation or state of uniform motion in a straight line of that laboratory from any other location, epoch, orientation or state of uniform motion in a straight line respectively.*

Note how strongly worded this is. We are asserting not only that one doesn't know where one is, but that one can't tell any difference between one's present location and any other. Similarly, for epoch, orientation and uniform motion in a straight line.

This postulate is thus a statement of invariance. In particular it declares invariance under *translations* from one location to another, or from one epoch to another, and invariance under *rotations* from one orientation to another, as well as invariance under *boosts* from one state of uniform motion in a straight line to another. These operations are all *transformations* which reflect an underlying symmetry of physics.

Einstein's First Postulate (the Principle of Relativity) may be regarded as a special case of this postulate. It corresponds to invariance under boosts. It should be noted though that there is nothing in the Invariance Postulate of which Galileo would not have approved. The special features of Einstein's Relativity theory stem from the details of the boost transformation — which we have not here specified, and need not at this stage.

All symmetries are related to the homogeneity and isotropy of space and time, together with our use of light to measure distance. If space and time were not homogeneous then we could not expect experiments performed at different locations and times to have the same outcome. However, the invariance postulate is a stronger statement, referring to physical processes and measurements rather than abstract geometry.

The postulate has some immediate consequences. Suppose, for example, that we have taken a measuring rod at rest and made marks on it one meter apart, according to our prescription for measuring (proper) length and using the definition of a meter. Now take that same rod at a later time and remeasure the distance between the marks. If they were not still one meter apart then the postulate would be violated; for nothing has changed except that the laboratory has been translated in time and our definition of length is independent of that. Similarly, if we rotate the rod and remeasure the distance between the marks, any discrepancy from one meter exactly would mean that there was something different between one orientation and the other. Since the orientation of the rod with respect to the walls of the laboratory can be well defined, this must imply something special about the way the laboratory is oriented in space. This would be contrary to the postulate. (Note that the independence of rod length on direction means that it can be used as a compass to trace out a circle as in Euclidean geometry.) Also, the rod length must be the same in any location as long as it is at rest. Furthermore, all of these equivalences must be true for the same rod at rest in any inertial frame.

What do these things mean? Recall that length measurement is to be done with standard clocks and light. The speed of the latter was defined to be a universal constant, c, in order to define distance units. We thus conclude that our definition was a sensible one and that, in terms of standard clocks and our length definition, the speed of light is indeed the same at every location, at every epoch, in every direction, and in every inertial frame. One small point is worth mentioning here. In defining length we have implicitly assumed that the light source used was at rest in the observer's inertial frame. These statements therefore only apply to light from a source at rest in an inertial frame. In the next chapter we shall return to the speed of light and make a strong statement applicable to moving light sources. Eventually we will even consider sources that are not moving uniformly in an inertial frame.

3.2 Nature of scientific experiments

Before proceeding, it is necessary to clarify a so far unstated assumption. It is basic to the entire scientific endeavor that experiments can be repeated and that their outcome depends only on the surrounding conditions and experiment design. To emphasize this assumption we elevate it to the status of an axiom.

Axiom 2 (Replicability)

The essential conditions in scientific experiments are replicable and solely determine the outcome.

Thus, different observers performing the same experiment, under the same set of circumstances, should get the same result, to within the accuracy of their apparatus and technique. This does not necessarily mean that *identical* results be obtained with *each* experimental run as some things, such as the toss of a coin or the throw of a dice, are inherently random. What it does mean is that if results do exhibit statistical fluctuations then repeating a large number of measurements should yield the same average and spread of results. Of course, if randomness is not involved, as in measurements of length and age, then different experimenters *in the same set of circumstances* should agree on the measured values. Since we cannot guarantee the absence of randomness beforehand, nor be absolutely certain regarding human error in the experimental process, it should be clear that in any scientific endeavor a single experiment is *never* enough. If we refer casually to a single measurement in this work it is only as an idealization, made with the knowledge derived from experience that measurements of that type are not random and are readily reproducible if conducted carefully.

Different results thus reveal differences in the subject of the experiment, or its surroundings inasmuch as they influence the experiment, and so enable the observer to explore, and analyze, his environment.

The axiom also precludes scientific study of things that are inherently not repeatable, such as claims of ghost sightings. Indeed, the axiom provides a very good test of what should be considered science and what is not.

3.3 Behavior of objects

With this background, it is possible to apply the invariance postulate and make statements about how certain things should behave in the laboratory. First though we need to know something about the kinds of things that we can perform experiments with. To this end we introduce an axiom governing the behavior of what may be termed classical particles, i.e. objects like billiard balls.

Axiom 3 (Classical particles)

1. *A classical point particle (and every point of an extended classical particle, or object) moves forward in time along a continuous trajectory.*

2. *A classical particle can only change its characteristic state when it is influenced by something external, such as if it collides with another particle. Otherwise, it retains its identity as it moves along.*

3. *A classical particle does not behave randomly.*

4. *Two (distinguishable) particles can not pass through each other and in a contact collision between two particles there is a repulsive interaction such that individual speeds are decreased in that direction in which motion is impeded by the other particle.*

The continuous trajectory is termed a *worldline*. The other statements summarize some common notions about classical particles. For example, a particle can't disappear at one point and reappear at another, with or without change.

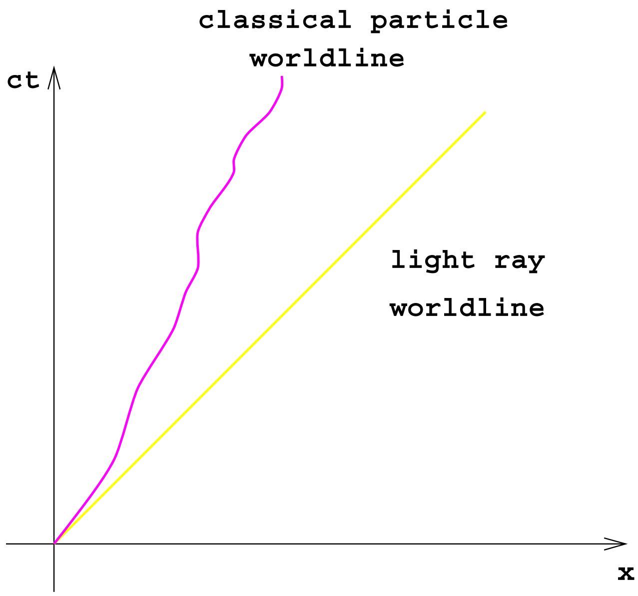

Figure 3.1: The worldline of a classical particle is a continuous trajectory on a spacetime diagram.

The axiom is by no means applicable to all particles. Quantum particles, such as photons and electrons, require their own axioms and separate treatment, though some properties may be shared with classical particles.

3.4 Spacetime diagrams

The worldline of a classical particle may be conveniently represented on a coordinate system plot such as that in fig. 2.3. We call the plot a *spacetime diagram*. Any event takes place at a particular point in space and time and is therefore represented by a point on the spacetime diagram. The worldline of a classical particle is a continuous line on the spacetime diagram. For the component of motion in one space dimension it may be represented as in fig. 3.1.

The slope of the worldline (i.e. the tangent at any point to the curve in fig. 3.1) is inversely related to the particle's instantaneous speed (in the direction of the space axis). We have

$$\begin{aligned}\text{slope} &= \lim_{\Delta x\to 0}\frac{c\Delta t}{\Delta x} = c\frac{dt}{dx} = \frac{c}{\frac{dx}{dt}} \\ &= c/v \\ &\equiv 1/\beta. \end{aligned} \tag{3.1}$$

Thus, an object at rest is simply a vertical line on the spacetime diagram.

Exercise 3.1 *Sketch a spacetime diagram showing the trajectory of a vehicle which starts at rest, then begins to move, slowly at first but then increasing up to a very high speed, at which it travels for a while before slowing down and coming to a halt. After parking for some time it then returns to its starting position through a similar sequence.*

Mathematically astute readers may question whether the limit in Eq. 3.1 always exists. Continuity of the trajectory is not sufficient, as an abrupt change in the rate of motion leads to non-unique limits at the point of change. For a free particle, the matter is covered by the considerations of the next section.

3.5 Newton's First Law

The Invariance Postulate and the classical particle axiom can be used to prove a result that is fundamental to classical physics. It is the first of Newton's three laws of motion on which he based his treatise on mechanics [11].

Proposition 1 (Uniform motion of free objects) *Viewed from an inertial reference frame, an object free of external influences continues in its state of rest or uniform motion in a straight line (i.e. Newton's First Law applies). Conversely, if objects free of external influence are observed to always behave in this manner then the reference frame is inertial.*

Before considering the proof of this proposition, we note that we can define *velocity* in the usual manner as the speed, together with the direction of motion. Thus velocity is a vector quantity and its instantaneous value is given by the derivative

$$\mathbf{v} = \frac{d\mathbf{x}}{dt} = \left(\frac{dx^1}{dt}, \frac{dx^2}{dt}, \frac{dx^3}{dt}\right) \tag{3.2}$$

Uniform motion in a straight line is just constant velocity and it is convenient to use this in formulating a proof.

The proof of the first part follows from translation invariance. Suppose we make a change of origin for our coordinate system so that $\mathbf{x}$ goes to $\mathbf{x}' = \mathbf{x} - \mathbf{X}$. Then, since $\mathbf{X}$ is constant we have that

$$\mathbf{v}' \equiv \frac{d\mathbf{x}'}{dt} = \frac{d\mathbf{x}}{dt} \equiv \mathbf{v} \tag{3.3}$$

This means that at any instant, the velocity is a quantity independent of the arbitrary choice of origin. Furthermore, it is a quantity which can be measured locally. However, the velocity can conceivably vary along the particle worldline (as in fig. 3.1) and therefore could be a function of the coordinates of that worldline:

$$\mathbf{v} = \mathbf{v}(x^{\mu}_{worldline}) \tag{3.4}$$

for $\mu = 0, 1, 2, 3$. Now actually, the fact that the particle moves along a worldline means that the time dependence is only implicit through the other coordinates, i.e.

$$\mathbf{v} = \mathbf{v}(\mathbf{x}(t)) = (v^1(\mathbf{x}(t)), v^2(\mathbf{x}(t)), v^3(\mathbf{x}(t))). \tag{3.5}$$

Furthermore, this reduces to (see problem 3)

$$v^i = v^i(x^i(t)) \,\mathrm{for}\, i = 1, 2, 3 \tag{3.6}$$

and this is the only dependence that the components of $\mathbf{v}$ have. Since we have found $\mathbf{v}$ to be independent of a (passive) translation of the origin, this means that the velocity is dependent on its intrinsic position in space. The object in question though is free of external influences and can therefore be regarded as in an enclosed laboratory. Hence the Invariance Postulate applies. The possibility that $\mathbf{v}$ depends on the location of the object in space is therefore denied because a measurement of velocity would provide information on the location of the laboratory, unless changes were random and equally likely at any position. However, classical particles do not behave randomly so it follows that $\mathbf{v}$ does not depend on $\mathbf{x}$. Since this is the only dependence $\mathbf{v}$ could have had we conclude that $\mathbf{v}(\mathbf{x}) = \text{constant}$.

Consider now the converse. We have a situation in which all objects free of external influence move at constant velocity. Suppose that the frame is not inertial, then we must accept that light rays would, in general, travel in curved paths. However, if we observe those light rays from an inertial frame, they necessarily travel in straight lines. Hence, if we observe from this same inertial frame, free objects moving between the same two points as the above light rays it follows that some free objects must travel curved paths. (This is simply because one can't straighten out the curved path of the light ray without bending the straight paths of the free objects.) But these objects were assumed free of external influence and thus, as established above, must travel in straight lines in any inertial frame. Hence we have arrived at a contradiction and must conclude that the frame is inertial.

Some useful results are immediate.

Corollary 1.1 (Inertial frames) *An inertial frame is one in which Newton's First Law is valid, i.e. one in which an object free of external influence continues in its state of rest or uniform motion in a straight line.*

Corollary 1.2 *Any inertial frame moves at a constant velocity with respect to any other inertial frame.*

The first of these results follows from the "if and only if" nature of the proposition and is a very common definition of an inertial frame. Observe though that the statement makes no sense as a definition because, until inertial frames have been defined, and a system of measurement prescribed, we can not meaningfully talk about uniform motion in a straight line.

The second corollary follows because an object free of external influence can be placed (at rest) at the origin (or any spatial coordinates) of an inertial frame. Since this object must move with constant velocity as observed in any other inertial frame, the frame itself must also move with constant velocity.

Review

What is a space translation? a time translation? a rotation? a boost?

What does the replicability of experiments mean?

What is a worldline?

What is the slope of a worldline on a spacetime diagram?

What are the properties of classical particles?

What is Newton's First Law?

Give two characterizations of inertial frames that follow from the uniform motion of free objects. Why are neither suitable as definitions of inertial frames?

Questions

1. Consider a vain runner holding a mirror in front of his face while running uniformly at close to the speed of light. Does the runner's image in the mirror appear any different than when he is standing still? [This problem played a significant role in Einstein's development of his relativity theory.]

2. We noted in the text that our methods for measuring distance led to a prescription for drawing a circle that is basically the same as Euclid's method using a compass. Is the geometry of the space dimensions in complete accord with the five basic postulates of Euclid? i.e.

   (a) A straight line can be drawn from any point to any point.

   (b) One can produce a finite straight line continuously in a straight line.

   (c) One can describe a circle with any center and radius.

   (d) All right angles are equal to one another.

   (e) If a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which the angles are less than two right angles.

   (Euclid defined a right angle via two intersecting straight lines; when adjacent angles between the lines are equal, the angles are right angles.)

3. A corner reflector utilizes two mirrors at right angles to each other to send light back in the direction from whence it came. Prove that the path travelled by the light satisfies Euclid's Fifth Postulate. (See the previous question. You will need to use the experimental fact that the

incident and reflected light rays make equal angles with a stationary reflecting surface.) What guarantee is there that the incident and reflected light rays are parallel (i.e. never meet) as asserted by Euclid?

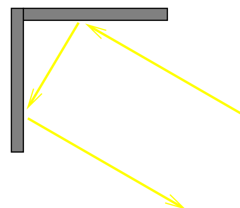

Figure 3.2: Corner reflector

4. Consider relaxing some of the axioms for classical particles. What consequences might changes have?

5. In the proof of Newton's First Law it was observed that velocity can be measured locally. Why is this important to the proof? Why is it important that velocity is invariant under a space translation? (Consider for example what happens to velocity under a rotation. Why is it permissible for velocity to change under a rotation but not under a space translation.)

6. Suppose we had a non-classical particle that changed its form at some position in the laboratory. Would this violate any of the fundamental symmetries of physics? If not, what consequences would those symmetries have for this process?

Problems

1. An observer has a set of cameras at regularly spaced intervals, together with synchronized stroboscopes which flash once every second. These enable him to record the position of a moving object. Consider the following sets of observations for three different circumstances.

ct	x	y
0	3.0	2.0
1	3.6	2.0
2	4.2	2.0
3	4.8	2.0
4	5.6	2.0
5	6.4	2.0
6	7.2	2.0

ct	x	y
0	3.0	2.0
1	3.6	2.0
2	4.2	2.0
3	4.8	2.0
4	4.8	2.6
5	4.8	3.2
6	4.8	3.8

ct	x	y
0	3.0	2.0
1	3.6	2.0
2	4.2	2.0
3	4.8	2.0
4	4.2	2.0
5	3.6	2.0
6	3.0	2.0

(a) Sketch the positions of each particle on separate xy plots and interpolate the trajectories.

(b) Making reasonable assumptions of uniformity, calculate the x and y components of velocity of each particle at various stages.

(c) Plot spacetime diagrams for both the x and y coordinates for each particle.

(d) Consider the effect of a shift in the x origin to $x = 3.0$. Are any of the velocity components changed?

(e) If these observations were repeated, the replicability of experiments and associated reproducibility of results would require either that the abrupt changes at $x = 4.8$ be reproduced or that a definite randomness in those changes be discovered. If the changes were found to be random, what would translation invariance require of those changes? If the changes are not random, can they be consistent with translation invariance for a free particle?

(f) If each particle is judged to be classical and free of external influence, could any of the above sets of observations have been made in a universe conforming to our axioms and Postulate of Invariance?

2. (a) Show that the angular velocity $\omega = \frac{d\theta}{dt}$ of a particle is independent of the direction chosen as the zero of angle. (The angle θ is that between a straight line through the object and a fixed point and some fixed reference direction. Changing the reference direction corresponds to a rotation about the fixed point.)

(b) Consider an object moving in a straight line at uniform speed v. Show that the angular velocity, ω, about any fixed point, is not a constant and that in fact, if r is the distance from this fixed point to the object then

$$\omega = \frac{v}{r}\cos\theta = \frac{vr_\perp}{r^2},$$

where θ is the angle from the direction of closest approach, and $r_\perp$ is the distance to the object at closest approach.

(c) Does the measurement of ω tell us anything about the orientation of an enclosed laboratory in which the measurement is conducted? Why? (Does ω really depend on θ? What would we have concluded if the object had been further away? Why does the result of part (a) not guarantee that ω is constant?)

(d) Show that the areas swept out by a line, joining any fixed point and the object, in equal time intervals, are the same for an object moving in a straight line at uniform speed. Are these areas distinct from the quantity $r^2\omega$?

(e) Is the quantity $r^2\omega$ independent of the location chosen as the zero of angle? Does it depend on any other variable?

(f) If we could keep r constant (i.e. circular motion) without destroying the isotropy of space, would ω necessarily be constant?

3. Consider carefully the procedure for constructing a spatial coordinate system.

   (a) Use the Invariance Postulate to prove that the x, y, z coordinates are independent of each other.

   (b) Show that if the i component of velocity, v^i, depends on the j coordinate then the coordinate x^i also depends on x^j. Hence, conclude that for $i \neq j$, $v^i \neq v^i(x^j)$.

Chapter 4

The Constancy of the Speed of Light

4.1 The speed of light

In the last chapter we noted that the Invariance Postulate required that the length of an object not change with orientation, position, or time; for nothing in the definition of length identifies the orientation or location of the laboratory in spacetime. This means that our adoption of a universal value of the speed of light, c, in our definition of length, was a sensible one — provided of course that experiment confirms the validity of the postulate. Thus invariance of c, incorporated in the definition of length, acquires meaning through the Invariance Postulate. Recalling that the definition of length only used sources at rest, what we in effect asserted (through combination of definition and postulate) is that the speed of light emitted by a source at rest in vacuum is the same for all inertial observers, independently of the location, epoch or orientation of the source.

Causal observation certainly suggests that this is true. Objects don't appear to change their length as they are moved about. However, we shall, in this chapter, consider some accurate experimental tests of the matter. That is, we shall examine some tests of this particular consequence of the Invariance Postulate.

However, none of this requires consideration of moving sources. If a light source at rest in one inertial frame emits light at speed c, what will its speed appear to be in another inertial frame — where the source will be moving at uniform velocity? There does not seem to be any way of answering this question based only on what we have considered so far.

We now introduce another postulate, often called Einstein's second postulate, which addresses the problem of moving sources. Einstein was inspired by consideration of electromagnetic phenomena and the internal consistency and elegance of Maxwell's theory of electromagnetic wave propagation. We shall simply make the postulate and directly examine the experimental evidence for it. In later chapters we pursue the profound consequences it has.

Postulate 2 (Light from moving sources) *The velocity of light in vacuum is the same for all inertial observers, independently of the uniform velocity of the source.*

When we consider things like the velocity of ripples on the surface of water this postulate seems extraordinary. For we know that if the water moves, as in a river, then the velocity with respect to the river bank depends on whether we are looking at a ripple moving upstream, or downstream, or across river. The essential reason why the postulate makes a different claim is that light requires no medium for its propagation. A vacuum being just that, there is no medium whose velocity need be considered.

The postulate can be combined with our other assertions about the speed of light.

Remark 2.1 (Constancy of c) *The speed of light in vacuum, in any inertial frame, is independent of the location, epoch, orientation, or uniform velocity of the source.*

We now examine some of the wealth of experimental evidence that has been accumulated on the constancy of c. Firstly, let us consider accurate experiments involving sources stationary with respect to the observer. Often the experiments are conducted in air. In a medium, light can be affected by the velocity of the medium (see Chapter 10), but if the air is stationary with respect to the apparatus that is not normally a problem.

4.2 Stationary sources

The earliest investigations of the speed of light were inspired by the unfounded assumption of physicists in the 18th century that light did move in a medium, which they called the ether. There was no real reason for these physicists to make this assumption because Maxwell's equations of electromagnetism predicted that electromagnetic waves could propagate in a vacuum. However, since no other wave phenomenon was known that did not need a medium for its propagation, physicists assumed that there must be a medium for light. The name "ether" came from the Ancient Greeks and was their solution to certain philosophical dilemmas posed in their perceptions of a complete void. The ether had to have extraordinary properties since it needed to be highly viscous, in order not to affect the motions of the planets, while at the same time it had to possess extremely high rigidity in order for the speed of light to be so great. It was believed that light propagation in the ether would be affected by the velocity of the source in the ether, just as water waves are affected by motion of the water. The goal of the early experiments was to detect variations of the speed of light due to the motion of the ether. Of course, nobody knew the state of motion of the ether but it seemed reasonable to suppose that Earth, with an orbital speed of 30 km/s, was moving through it at roughly that same speed. At least it seemed reasonable that the changing direction of the Earth's velocity as it moved in its orbit should be reflected in variations of that order in its speed through the ether.

We shall not bother to address the details of the predictions of the ether theory but shall rather analyze all experiments from a modern point of view, in keeping with the assumptions and postulates we have considered so far. The ether theory results follow as special cases and are not of particular interest, since the theory fails so miserably.

Though we have defined the speed of light for stationary sources to have a fixed value, our definition of length depends on that and any variation in that length could be reinterpreted as due to a variation in c. From the definition of length, such a variation must have the form

$$\frac{\Delta l}{l} = \frac{\Delta c}{c}. \tag{4.1}$$

Since any variation in length could be expected to be greater for longer objects, we find it more convenient in this chapter to simply rephrase things in terms of a variation, Δc, in c.

The Michelson-Morley experiment

Albert Michelson was an American physicist who developed an outstanding reputation for his ingenious and precise experimental work. He was the first American to win the Nobel Prize in physics (in 1907). While still young, he developed a new instrument of unprecedented sensitivity, called an interferometer, to search for variations in the speed of light due to motion of Earth in the ether.

A schematic diagram of Michelson's interferometer is shown in fig. 4.1 Light from a source S is directed on a partially silvered mirror P, at 45° to the source. Some of the light passes through, where it reaches mirror M_1 and is reflected back; it then is reflected by the partially silvered surface into the viewing telescope. The rest of the light is reflected at P towards mirror M_2, where it is reflected back and passes through P to the telescope. The two light beams interfere with each other, leading to a fringe pattern, visible in the telescope. The fringe pattern arises from phase differences between the two beams. (Note that one is actually seeing light from slightly off-axis as well as straight through.) If the beams are in phase, one gets constructive interference and a bright fringe; if the beams are out of phase then there is destructive interference leading to a dark region between the successive bright fringes.

The round-trip time in arm 1 of the interferometer is (using the definition of length)

$$t_1^0 = 2l_1/c_1 \tag{4.2}$$

and that in arm 2 is

$$t_2^0 = 2l_2/c_2. \tag{4.3}$$

Here, l_1 and l_2 are the respective lengths, and c_1 and c_2 are the respective average round-trip speeds of light in the directions of the two respective arms. The time difference,

$$\Delta t^0 = t_1^0 - t_2^0 \tag{4.4}$$

determines whether the two beams are in phase or out of phase.

The two arms are at right angles to each other so if there really is a dependence of the round-trip speed of light on direction in space, then after

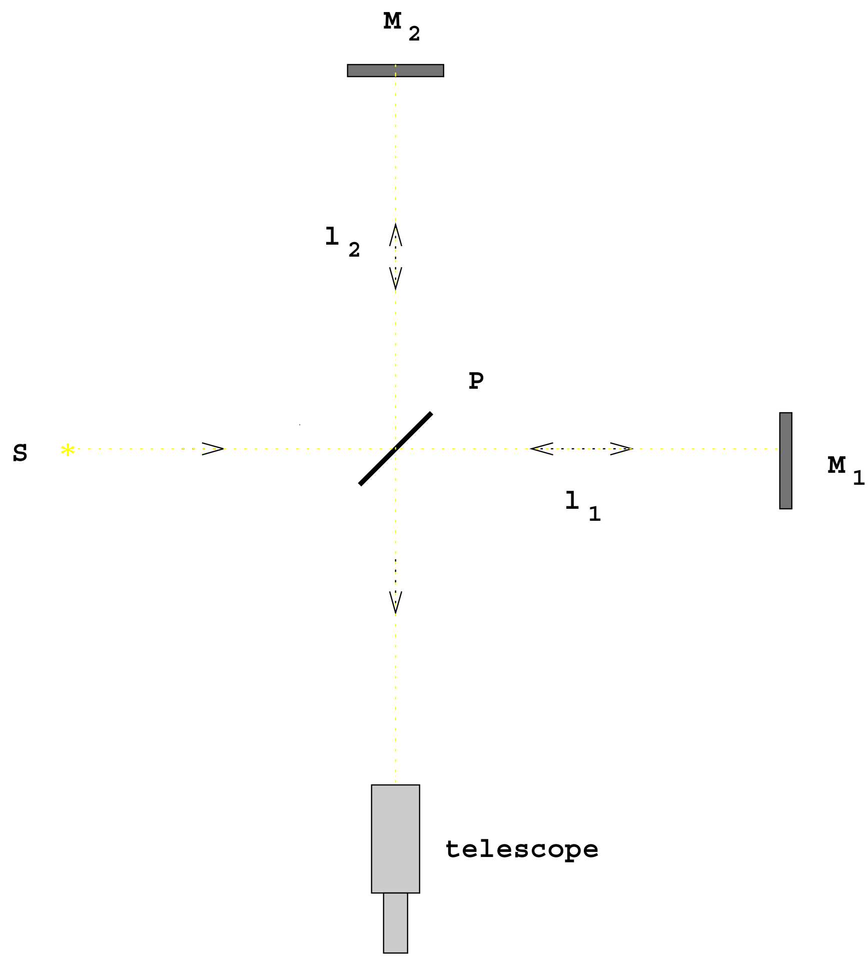

Figure 4.1: Schematic of Michelson's interferometer

a rotation of the apparatus by 90° the velocities in the two arms will be reversed. Then, the time in arm 1 will be

$$t_1^{90} = 2l_1/c_2 \tag{4.5}$$

and that in arm 2 will be

$$t_2^{90} = 2l_2/c_1, \tag{4.6}$$

leading to a difference

$$\Delta t^{90} = t_1^{90} - t_2^{90}. \tag{4.7}$$

The difference between these is

$$\begin{aligned} \Delta t &= \Delta t^0 - \Delta t^{90} \\ &= 2(l_1 + l_2)\left(\frac{c_2 - c_1}{c_1 c_2}\right), \end{aligned} \tag{4.8}$$

and if non-zero, leads to a shift in the fringe pattern. For

$$\Delta c \equiv c_2 - c_1 \ll c \approx c_1 \approx c_2 \tag{4.9}$$

we have

$$\Delta t = 2(l_1 + l_2)\Delta c/c^2 \tag{4.10}$$

and the corresponding difference in optical path lengths is

$$\Delta d = c\Delta t = 2(l_1 + l_2)\Delta c/c. \tag{4.11}$$

Each multiple that this is of the wavelength of light, λ, corresponds to a shift by one fringe. Hence the number of fringes by which the fringe pattern shifts is

$$\delta = \frac{\Delta d}{\lambda} = \frac{2(l_1 + l_2)}{\lambda}\left(\frac{\Delta c}{c}\right). \tag{4.12}$$

Michelson originally performed the experiment in 1881, with null results, but the errors turned out to be of the same order as the effect he expected to find and he repeated the experiment in 1887 with Morley. In the Michelson-Morley experiment [12] the two arms were chosen to be of equal length and a compensating piece of glass was even inserted in arm 2 so that light in both arms travelled through the same amount of glass on its way to the telescope. Considerable effort was expended to make the experiment more accurate, including mounting the whole apparatus on a slab of marble floating in mercury to reduce vibrations. (Horse drawn traffic in the street outside was enough to cause troublesome vibrations and many experimental runs were performed in the early hours of the morning to reduce these.)

After statistical averaging, Michelson and Morley claimed that any fringe shift was no more than 0.01 fringes. With

$$l \equiv l_1 = l_2 = 11\,\mathrm{m} \tag{4.13}$$

and sodium light of wavelength

$$\lambda = 589\,\mathrm{nm} \tag{4.14}$$

this corresponds to

$$\Delta c < 4\,\mathrm{cm.s}^{-1}. \tag{4.15}$$

Clearly then, any dependence of the speed of light on orientation must be very small.

In the ether theory, the maximum difference in round-trip light speeds can be shown to be (see Problem 2)

$$\Delta c = v^2/2c \tag{4.16}$$

where v is the speed of the supposed ether. The experiment thus lead to the conclusion that any movement with respect to an ether had to be less than 5 $\mathrm{km.s}^{-1}$. That this is much larger than the limit on Δc is due to the manner in which Δc arises in the ether theory.

Several researchers have repeated the Michelson-Morley experiment with the most sensitive repeat by Joos [13], yielding about an order of magnitude improvement in the limit on Δc. An indication of the difficulties involved may be seen by noting that one such repetition did find an effect but it was not correlated in any way with the orientation of Earth. It is nowadays believed [14] that the observed fringe shift was due to a 1° temperature variation across the room, which had snow packed against the exterior of one side! (The speed of light in air is not quite the same as that in vacuum and a density variation due to the temperature gradient would lead to minute variations in the speed of light in air.)

It is also worth noting that the Michelson-Morley experiment has been performed with extraterrestial light sources, including both sunlight and starlight.

The Lorentz-Fitzgerald contraction hypothesis

In 1892, Lorentz and Fitzgerald independently proposed an explanation for the null result of the Michelson-Morley experiment. They suggested that there was indeed a significant difference in the speed of light in the two arms but that one of the arms actually contracted in just such a manner as to cancel the effect of the different speeds.

Suppose that the length of arm 1 was actually

$$l_1' = l_1/\gamma \tag{4.17}$$

and that after rotation by 90° arm 2 gets contracted by the same factor. Then

$$\begin{aligned} \Delta t &= (2l_1/\gamma c_1 - 2l_2/c_2) - (2l_1/c_2 - 2l_2/\gamma c_1) \\ &= 2(l_1 + l_2)\frac{(c_2 - \gamma c_1)}{\gamma c_1 c_2}. \end{aligned} \tag{4.18}$$

If $\gamma = 1$ then there is no contraction factor and the previous results are recovered. If however, $\gamma = c_2/c_1$ then a null result is obtained.

As bizarre as this hypothesis may sound, it can not be easily dismissed. The matter was finally addressed in a classic experiment by Kennedy and Thorndike in 1932.

The Kennedy-Thorndike experiment

The essence of the Kennedy-Thorndike experiment [15] is not to rotate the Michelson interferometer, as in the Michelson-Morley experiment, but rather to leave it in a fixed orientation with respect to the laboratory and to look for diurnal and seasonal variations as that laboratory rotates due to Earth's spin and orbital motion.

Consider Earth in orientation A. Then the time difference between the two arms of the interferometer is (according to the Lorentz-Fitzgerald hypothesis)

$$\begin{aligned}\Delta t_A &= \frac{2l_1}{\gamma^A c_1^A} - \frac{2l_2}{c_2^A} \\ &= \frac{2(l_1 - l_2)}{c_2^A}\end{aligned} \tag{4.19}$$

whereas in orientation B it is

$$\Delta t_B = \frac{2(l_1 - l_2)}{c_2^B}. \tag{4.20}$$

Thus, for $l_1 \neq l_2$ a fringe shift should be observed, due to the different speeds of light.

The Kennedy-Thorndike experiment used $l_1 - l_2 = 16\,\mathrm{cm}$, which is at the limit of what could be tolerated while maintaining coherence of the light in both arms. No such diurnal or seasonal variation was seen. The limit set on $c^A - c^B$ was about $3\,\mathrm{m.s^{-1}}$.

Note that this experiment also shows that there is no variation in c with position of the Earth in its orbit, or with time. (Actual run-times varied from eight days to a month.) Thus the experiment provides additional information on c and the fact that it sets a less stringent limit on Δc than the Michelson-Morley experiment does not undermine its importance.

Modern laser experiments

Lasers permit a more accurate version of the Michelson-Morley experiment to be undertaken. A laser operates by the process of stimulated emission whereby a photon of the right frequency causes an electron in an excited atomic state to make a transition to a less excited state, with the emission of a second photon of the same phase and wavelength as the original photon. If these photons now cause further stimulated emissions then the amplitude of the light in phase with the original photon increases. The trick is to arrange for sufficient electrons in excited states so that the rate of stimulated emission exceeds the rate at which photons are absorbed by the reverse process in which an electron is promoted to the excited state. Also, one creates a resonant cavity in which the myriad photons will be reflected from the ends and still be in phase with others in the middle of the cavity. See fig. 4.2.

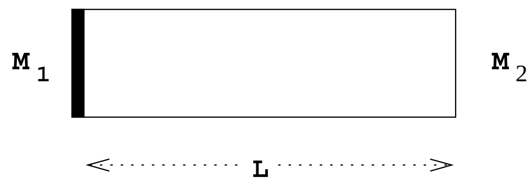

Figure 4.2: Resonant cavity in a laser. The mirror M_2 is partially silvered allowing a fraction of the laser light to escape.

The partially silvered mirror M_2 at one end reflects most light back but allows a little to escape, forming the laser beam. The resonant frequency of the cavity is

$$\nu = c/\lambda \tag{4.21}$$

where, for resonance,

$$\lambda = 2L/n, \; n \text{ an integer.} \tag{4.22}$$

Now suppose that one has two lasers, oriented as in fig. 4.3, much like the Michelson-Morley arrangement. If the speed of light is different for the two lasers in their respective orientations then there would be a slight difference in their frequency and the superposition of the two light beams at the half-silvered mirror would lead to a beat frequency given by

$$\nu^0_{beat} = \nu^0_B - \nu^0_A = \frac{n}{2}\left(\frac{c_B}{L_B} - \frac{c_A}{L_A}\right). \tag{4.23}$$

Now rotate the apparatus through 90°, leading to

$$\nu^{90}_{beat} = \nu^{90}_B - \nu^{90}_A = \frac{n}{2}\left(\frac{c_A}{L_B} - \frac{c_B}{L_A}\right). \tag{4.24}$$

Thus

$$\Delta\nu_{beat} = \nu^0_{beat} - \nu^{90}_{beat} = \frac{n\Delta c}{2}\left(\frac{1}{L_B} + \frac{1}{L_A}\right) \tag{4.25}$$

where $\Delta c = c_B - c_A$. For $L_A \approx L_B = L$ this becomes

$$\Delta\nu_{beat} = \left(\frac{n}{L}\right)\Delta c. \tag{4.26}$$

Noting that the average frequency of the laser light is

$$\overline{\nu} = nc/2L \tag{4.27}$$

we obtain

$$\frac{\Delta\nu_{beat}}{\overline{\nu}} = 2\frac{\Delta c}{c}. \tag{4.28}$$

In an experiment by Jaseja, Javen, Murray and Townes [16] in 1964, laser light of frequency $\overline{\nu} = 3 \times 10^{14}$ Hz (in the infra-red) was used. It being possible to maintain frequency stability of the lasers used to within just 20 Hz, this experiment was able to conclude that $\Delta\nu_{beat}$ was less than 3 kHz, leading to

$$\Delta c < 10^{-3}\,\mathrm{m.s}^{-1}. \tag{4.29}$$

This experiment is so sensitive that actual changes in length due to the "magnetostriction" effect of Earth's magnetic field on the apparatus can be readily observed. The null result comes from repeating the experiment over several hours and showing that length changes have the same magnitude, independently of Earth's orientation. This unfortunately limited the accuracy of the experiment to only a few times better than that achieved by Joos.

In a refinement of this experiment, Brillet and Hall [17], in 1979, used a servo mechanism to electronically lock the frequency of a laser to the resonance conditions of a Fabry-Perot etalon. This apparatus was mounted on a rotating granite slab and the laser frequency was compared with that of a fixed reference laser. With this arrangement, they were able to obtain a 4000-fold improvement on the experiment of Jaseja et al., showing that any variation in the round-trip speed of light with orientation must be such that

$$\Delta c < 0.5 \times 10^{-6}\,\mathrm{m.s}^{-1}. \tag{4.30}$$

4.3 Moving sources

While the above results on the constancy of c were unexpected by physicists of the 19th century, they can be understood as simply reflecting the Invariance Postulate, and might well have been anticipated if physicists had not mistakenly believed in the necessity of a medium for the propagation of light. Einstein's postulate on moving sources is more far-reaching though; in fact its ramifications are simply stunning. Unfortunately, it is not easy to verify directly. Nevertheless, convincing experimental evidence of its validity exists.

Photons from pion decay

The neutral pion, π^0, is an unstable subatomic particle with a mean lifetime of about 8×10^{-17} s [18]. It's principal decay mode is into a pair of photons:

$$\pi^0 \to \gamma\gamma. \tag{4.31}$$

These photons are not in the visible region of the spectrum but are gamma rays. Nevertheless, there is no reason to believe that their speed should differ from that of electromagnetic radiation in any other region of the spectrum. In fact there is good evidence, particularly from astronomical observations of distant sources, that all forms of light travel at the same rate in vacuum.

In an experiment [19] conducted at the CERN particle accelerator laboratory in 1964, the speed of these photons was measured directly from their time of flight. At CERN, protons can be accelerated to very high speeds in an evacuated beam pipe. When they strike a target, pions are produced. These pions themselves have very high speed and thus constitute a moving light source when they decay. In the experiment, the speed of the π^0s was determined to be $0.99975c$, i.e. very close to the speed of light. Despite this very high speed, the very short lifetime of the π^0 before it decays means that it travels at most a few μm. Thus the distance travelled by the photons before reaching a detector can be determined with good accuracy. From the time of flight, it was found that the photon velocity was

$$v_{photon} = 2.9977 \pm 0.0004 \times 10^8 \,\mathrm{m.s}^{-1} \tag{4.32}$$

Within experimental error, this is the same as the speed of light emitted by a source at rest.

If one were to hypothesize that the speed of light from a moving source acquired a component of the source velocity according to

$$c_{moving} = c_{rest} + k v_{source} \tag{4.33}$$

where k is a constant (that would be 1 in the ether theory) then the experiment sets the limit

$$k < 10^{-4}. \tag{4.34}$$

Binary stars

In 1913 de Sitter suggested that observations of binary stars could place stringent limits on any dependence of c on source velocity. Binary stars consist of a pair of stars orbiting a common central point. Observations with telescopes reveal that the orbital motion conforms well to the same pattern observed for planets in our solar system and can be accurately described with the same theories. From the rate at which the stars appear to be orbiting each other, and our knowledge of astronomical distance scales, we can deduce that the orbital speeds of these stars can be quite high.

Suppose that the speed of light depended on source velocity. Then, since at various stages in its orbit one of these stars has a varying velocity with respect to an Earth based observer, the speed at which light is emitted towards Earth should vary according as to where the star was in its orbit. However, even if the variation in light speed was quite small, the enormous distances to the nearest stars means that there would be a significant difference in the time taken for light to reach Earth. Consequently, light could be seen from a star in one position before light was received from the same star in an earlier position! Thus all sorts of strange effects could be observed, including ghost images and highly unusual orbital eccentricities.

No such effects are seen. However, there is a problem with concluding from this that the speed of light does not depend on source velocity. It turns out that the interstellar regions are not a perfect vacuum but contain minute amounts of gas. When light passes through any medium it gets continually absorbed and reemitted by the atoms in that medium. Thus, any dependence of the light on the velocity of the orginal source would be lost and replaced by whatever dependence existed on the velocity of the new source, i.e. the atoms in the medium. The distance over which this happens is called the extinction length. For air at normal atmospheric density this extinction length is only 0.1 mm for visible light! For the interstellar gas in the galactic disk of our Milky Way galaxy it is about 2 light-years. However, this is still less than the distance to the nearest star.

Fortunately, the extinction length is larger for other regions of the electromagnetic spectrum and for certain x-rays it is about 60,000 light-years. A handful of binary star systems are known, within this range, which contain suitable x-ray sources that even pulse regularly. From these one can conclude

[20] that if the speed of light is related to source velocity as in Eq. 4.33, then

$$k < 2 \times 10^{-9}. \tag{4.35}$$

Einstein's postulate on the constancy of c is thus verified to high accuracy.

Review

What essential difference exists between light and water waves and can be held responsible for the velocity of light being independent of source velocity while the velocity of water waves depends on source velocity.

What limit did the Michelson-Morley experiment set on variations of the round-trip speed of light with orientation? What is the best modern limit on this?

What is Earth's orbital speed?

What experiments provide *direct* support for Einstein's second postulate (i.e. independence of source velocity)?

Questions

1. Does it make any sense to postulate that the velocity of light in vacuum is the same for all inertial observers, including those that might be travelling faster than the speed of light in another inertial observer's frame?

2. Why did de Sitter suggest looking at binary stars for dependence of c on source velocity instead of just observing the moons of planets in our solar system?

3. It has frequently been claimed that Einstein's postulate on the constancy of c is inessential. Is this really the case? Can you suggest alternative postulates that might enable *derivation* of the constancy of c? (Readers needing a firmer grasp of the logic involved in this question may wish to compare this situation with attempts to prove Euclid's Fifth Postulate. In summary, all claimed proofs have been shown to hinge on some other unproven assumption and thus the net result is

equivalent to the Fifth Postulate, which nowadays is clearly recognized as an essential element of Euclidean geometry.)

Problems

1. What change in length of a 1 meter long measuring rod would be implied if there were a variation of $4\,\mathrm{cm.s}^{-1}$ in the speed of light as the direction was changed? (This is the limit imposed by the Michelson-Morley experiment.) What is the maximum change allowed by the Brillet-Hall experiment? Compare these changes with typical atomic dimensions: $10^{-9}\,\mathrm{m}$ for the spacing between atoms in crytals, $10^{-10}\,\mathrm{m}$ for the diameter of a hydrogen atom, $10^{-15}\,\mathrm{m}$ for the diameter of a hydrogen nucleus.

2. In Joos' 1930 repetition of the Michelson-Morley experiment, $\delta < 0.002$, $l = 21$ m and Na light was used. What limit does this impose on Δc?

3. A model speedboat race is held on a large river. To avoid collisions and concentrate on raw power the race organizers set up a dual course with two arms. Course A requires the boat to race 100 m directly upstream, round a buoy and return to the starting point. Course B requires the boat to race to a buoy 100 m directly across stream, round it and return to the starting point. Boats are assigned to either course A or B and the first one back is declared the winner.

   (a) Is the race fair?

   (b) If not, which course provides the advantage? Give a detailed analysis of this problem. (You may assume that the current in the river is uniform over the whole course, and assume also that the velocities of the boats and the current flow are small, comparable in magnitude, and add vectorially.)

4. In the ether theory, the velocity of light is c with respect to the ether and the velocity of light seen by an observer is found by adding the velocity of the ether vectorially to the velocity of light in the ether.

   (a) Show that the speed of light upstream (in the ether) is $c - v$, that downstream is $c + v$, and across stream it is $\sqrt{c^2 - v^2}$.

(b) Show that the average round trip speed of light in the direction of ether drift is $c_1 = c - v^2/c$.

(c) Use the binomial theorem to show that the difference between the average round trip speeds, parallel and perpendicular to the direction of ether drift, is

$$\Delta c \approx v^2/2c.$$

as stated in Eq. 4.16.

5. Show that in the ether theory (see the previous problem) the Lorentz-Fitzgerald contraction factor γ is

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

where v is the ether drift velocity.

6. Dr. Zapffe is an engineer, and an intelligent man, who believes that because the Earth is enveloped by a magnetosphere, the null result of the Michelson-Morley experiment says nothing about the absence of an ether as a medium for the propagation of light. In his view, the magnetosphere could simply have caused the ether to be dragged along at Earth's surface and by performing the experiment outside of the magnetosphere the ether wind could be evident. Therefore, he argues, the foundations of Einstein's relativity theory (i.e. the null result of the Michelson-Morley experiment) are unsound and the theory cannot be trusted. Dr. Zapffe has become frustrated at the apparent refusal of physicists to accept his concerns. Write a (polite) response to Dr. Zapffe explaining either why you agree with him or why you don't. [Your letter should be cohesive and persuasive and should display both accuracy and knowledge of the relevant physics. At a minimum, you should include some discussion of the theoretical and experimental foundations of special relativity.]

7. Consider a star in a circular orbit about the common center of a binary system. Let the radius of the orbit be r and the orbital speed be v. Let this star be observed from Earth, a distance $D \gg r$ from the center of the binary system, which is at rest with respect to Earth. For uniform v, the speed in the direction of an Earth based observer

varies as $\cos\omega t'$ where t' is the time according to a clock in the binary system, synchronized with clocks on Earth, and $\omega = 2\pi/T$, where T is the orbital period. Suppose that the speed of light depends on source velocity according to Eq. 4.33.

(a) Show that the time of arrival at Earth of light from this star is

$$t = t' + \frac{(D - r\sin\omega t')}{(c + kv\cos\omega t')}$$

(b) Show that for $v \ll c$ this reduces to

$$t \approx t' + \frac{D}{c} - \frac{r}{c}\sin\omega t' - \frac{kDv}{c^2}\cos\omega t'$$

and that this may be written as

$$t = t' + \frac{D}{c} - \frac{V}{c}\sin(\omega t' + \phi)$$

where $V = r\sec\phi$ and $\phi = \tan^{-1}\frac{kD\omega}{c}$.

(c) Sketch t vs. t' and observe that if the slope should ever become negative then light from the star in different positions in its orbit would arrive at Earth at the same time. Thus ghost images of the star would appear.

(d) Show that the condition for ghost images not to appear is that

$$k < k_g \equiv Tc^2/2\pi Dv.$$

(e) The binary x-ray source Cen-X3 at a distance of about 25,000 ly from Earth has an orbital speed of 415 km/s and a period of 2.09 days. What limit does the absence of ghost images of this source place on k?

8. Consider a binary star system whose orbital plane lies in the line of sight from Earth so that eclipses will occur. For simplicity, consider a small star orbiting just above the surface of a larger one. Then eclipses will begin at $t' = \pi/\omega$ and end at $t' = 2\pi/\omega$.

(a) Find the times, as seen by an observer on Earth, at which the eclipse will begin and end and show that the average (or mid-eclipse time) is

$$t_e = \frac{3\pi}{2\omega} + \frac{D}{c}.$$

independent of k (see Eq. 4.33).

(b) As the star orbits, the frequency of the radiation seen at Earth will be doppler shifted (see Chapter 7). The amount of the doppler shift varies in proportion to the derivative dt/dt'. Use the result from the previous problem to show that

$$\frac{dt}{dt'} = 1 - \frac{V}{c}\cos(\omega t' + \phi)$$

(c) If k were zero then mid-eclipse corresponds to $dt/dt' = 1$. Show that in general the derivative has this value for

$$t' = (3\pi/2 - \phi)/\omega$$

and observe that although the star is in eclipse here that this is just the mid-point of the maximum and minimum in the derivative and so can be determined by observations.

(d) Find the corresponding time as seen by an Earth based observer (see the previous problem) and show that there is a discrepancy between this and t_e determined above, given by ϕ/ω.

(e) Show that the absence of a discrepancy in the times determined by these two methods implies that

$$k < k_p = \phi c T/2\pi D$$

(f) The binary x-ray source Her X-1 is located about 20,000 ly from Earth. It has a period of 1.70 days and the discrepancy in the orbital phase determined from mid-eclipse and Doppler shifts is less than 0.06 radians. What limit on k does this impose?

Literature Project: The Michelson-Morley Experiment

Locate a copy of the original paper by Michelson and Morley: *The American Journal of Science*, third series, vol. 34, 1887, p.333. State where you found it and, briefly, how you found it. If you can't find it, state what steps you took and what further (unreasonable or expensive) steps you could take to obtain a copy and then see your instructor who should be able to provide a copy. [This information should be recorded as part of your assignment.]

Read the paper!

1. What, precisely, was the experimental hypothesis they set out to investigate?

2. The experiment was a refinement of an earlier experiment by Michelson. It is remarked that the earlier theoretical analysis was incorrect. What did the earlier analysis overlook? This paper itself contains an error in its discussion of stellar aberration as seen through a water-filled telescope — an experiment first performed by Airy with a null result. Michelson and Morley claim that the amount of aberration observed with a water filled telescope should be "four-thirds of its true value" — which of course Airy's experiment found was not the case. The 4/3 is a rough value for the refractive index of water and should actually be squared in the ether theory. Comment on the relative significance of the two errors.

3. In the ether theory, and for an interferometer with arms of equal length L oriented at an angle θ to the ether wind (where $\theta = 0^\circ$ corresponds to the case where the wind is parallel to one arm), the time difference for light travelling the two arms is given to a good approximation by

$$\Delta t(\theta) = \frac{v^2 L}{c^3} \cos 2\theta$$

where v is the supposed speed of the ether wind. Discuss briefly what this means in terms of what should be observed if the instrument is continuously rotated. (Note that the time difference between the two orientations of the apparatus discussed in the text corresponds to the difference between the maxima and minima of $\Delta t(\theta)$.)

4. What steps were taken in this experiment to get a more accurate result? Use a theoretical analysis where appropriate to demonstrate the greater sensitivity of the new apparatus. How big a fringe shift was expected in the new experiment?

5. Study the tables of data and the figures. What do the columns of data and the means represent? How is this data related to what is plotted in the figures? Comment on the level of explanation in the paper. Is it acceptable do you think?

6. How small a phase shift could Michelson and Morley detect? What do they claim? How reasonable is that? How do you think they achieved their claimed accuracy? What limit on the ether wind does the data of Michelson and Morley imply? (Be careful not to swallow the claims of Michelson and Morley without verifying for yourself.)

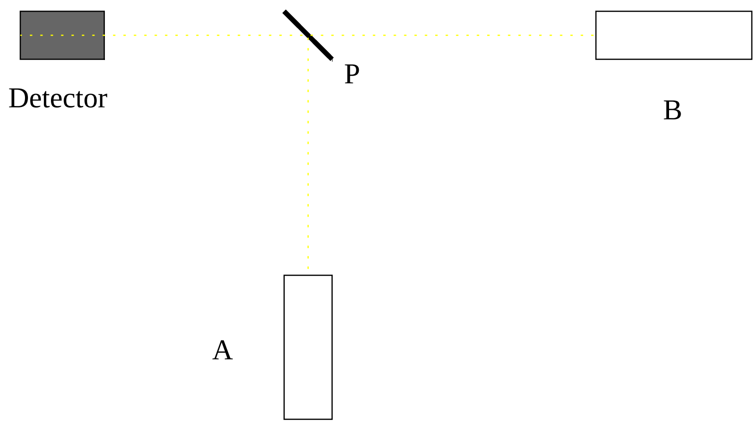

Figure 4.3: Laser arrangement for testing the isotropy of the speed of light. The beams from two lasers, A and B, are combined by the partially silvered mirror P and observed at the detector.

Chapter 5

Simultaneity

5.1 Relativity of simultaneity

Of all of the ramifications of the constancy of the speed of light, few are more startling than its implications for the concept of simultaneity. We say that *two events are simultaneous if, according to our system of synchronized clocks, they take place at the same time*. This seemingly innocuous statement turns out to be quite a can of worms!

Consider the following situation, often referred to as Einstein's Train Paradox. A high speed train is moving along some tracks when a pair of lightning flashes strike the ground, one at the front of the train and one at the back. See fig. 5.1. According to an observer on the ground, O_g, the two strikes are simultaneous. Thus, if this observer is actually located midway between the front and back of the train at the time of the strikes, he will see both at precisely the same time (since the light travels the same distance from each, and at the same speed).

Now consider what an observer, O_t, who is on the train, midway between the front and back, will see. The light radiating from each lightning flash travels in a single wavefront. Since both wavefronts reach O_g at the same time, after having started from the front and back of the train, and since O_t is moving into the light from the front and away from the light from the back, O_t *must* see the light from the front arrive first. There is no doubt that this is what will occur if the light is travelling at a uniform rate (i.e. independent of location) as seen by O_g. The question is, what does this mean? To a physicist of the 19th century, the answer would simply be that the light from

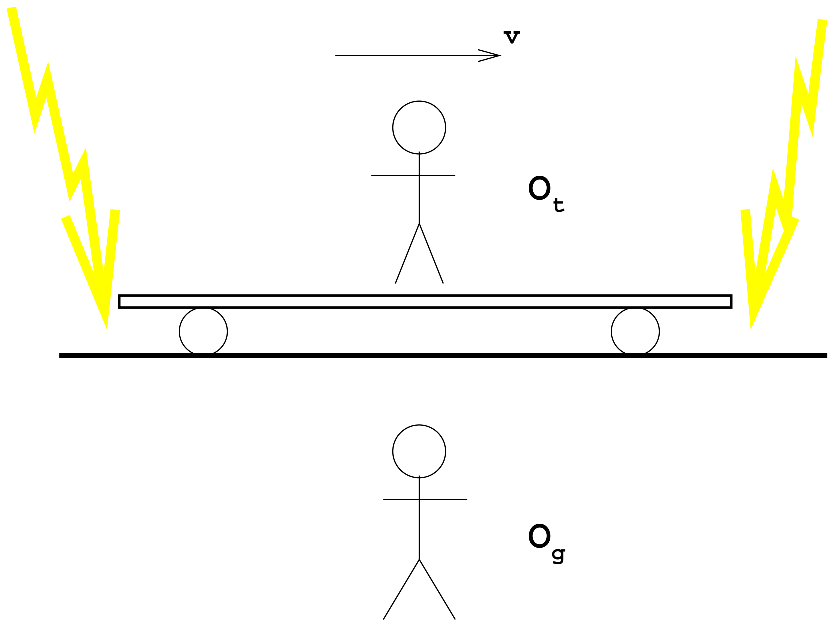

Figure 5.1: Lightning strikes the ground at both ends of a moving train. According to a ground based observer, O_g, the strikes are simultaneous.

the front of the train travels at a greater velocity relative to O_t than light from the back. But we have seen in the last chapter that light travels at the same speed for all inertial observers, independently of the velocity of the source. (Both observers in this case can be considered to be in restricted inertial frames and so Einstein's Second Postulate does apply to the speed of light in a direction parallel to the tracks. For those not happy with this, a similar situation can be readily constructed with light beacons in the furthest reaches of outer space.) Thus, even though the lightning strike represents a source at rest in the ground frame, the train-based observer must see the light from both front and rear strikes approach him at the same rate. The only possible conclusion then is that the front strike actually occurred first!

The complexity of the phenomenon of lightning cannot be blamed for this. It could happen that the lightning could cause damage to the front and back of the train. O_g would still see the damage appear simultaneously while O_t would see the damage at the front appear first. Furthermore, O_g must agree that O_t will see things this way because we could arrange to place atomic clocks at the front and back of the train with large digital readouts (visible to ground-based observers) that could be stopped by the lightning strikes. O_t can synchronize these clocks according to the standard procedure in his frame. O_g will note that according to the train clocks the front strike occurred first. (The stopped rear clock will read a later time than the stopped front one.) He will though simply conclude that the clocks were not correctly synchronized (or at least not synchronized according to his procedure for doing so).

Thus we are forced to conclude that the notion of simultaneous events depends on the reference frame: events simultaneous in one inertial frame will not be simultaneous in another inertial frame. This quirk lies at the heart of many of the seeming perplexities in the theory of relativity and many of the attempts to debunk the theory can be dismissed for failure to appreciate this aspect.

5.2 Representation on a spacetime diagram

If one considers the synchronization procedure on a spacetime diagram, one readily sees that the lines of simultaneity for a moving observer will be tilted to those for a stationary observer (where stationary and moving refer to the frame of the axes in the diagram). See fig. 5.2. Let a light source be placed

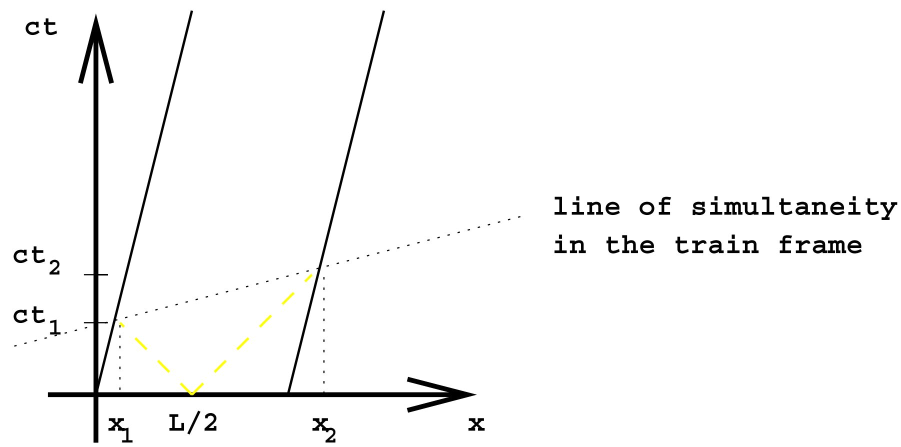

Figure 5.2: Spacetime diagram for an object of length L moving to the right. Worldlines of front and back are shown as light solid lines. A pulse of light (dotted line) is emitted towards front and back from the center of the train. In the train frame it reaches the ends simultaneously.

midway on a train of proper length L_0. Both train-based and ground-based observers will agree on the midpoint of the train, just as they agree on the endpoints. The train-based observer will see the light reach both ends of the train simultaneously. However, the ground-based observer must discover that the light reaches the back of the train at t_1 before it reaches the front of the train at t_2. This is because the back of the train is moving towards the light while the front is moving away from it. Thus the line of simultaneity for the train-based observer appears tilted in a spacetime diagram for the ground-based observer.

If the train is moving in the opposite direction, then the lines of simultaneity will be tilted the opposite way.

We can use a diagram such as this to assist us in determining just how great the difference in (coordinate) time is for events that are simultaneous in one frame but not in that frame in which the (coordinate) time is measured.

In the ground-based observer's frame, let the train length be L and let the light leave $L/2$ at $t = 0$. If the train speed is v then in time t_1 the back of the train travelled a distance vt_1 (reaching x_1) while light travelled ct_1 to get there. Similarly, in time t_2 the front of the train travels vt_2 (reaching x_2)

while the light travels ct_2 to reach it. Thus

$$(v+c)t_1 = L/2 = (c-v)t_2. \tag{5.1}$$

A few lines of algebra shows that

$$t_2 - t_1 = \frac{v}{c^2 - v^2} L \tag{5.2}$$

The deviation from simultaneity thus increases with the speed, v, and depends on the spatial separation, L, of the events (i.e. light reaching front and back of the train).

Exercise 5.1 *Verify the expression for $t_2 - t_1$ and show that*

$$t_1 + t_2 = \frac{c}{c^2 - v^2} L \tag{5.3}$$

(We shall use this result later.)

5.3 Length measurements

The relativity of simultaneity has major implications for length measurements. If we are measuring the proper length, i.e. that of an object at rest in our frame, then there is no problem. However, the measurement of the length of a moving object requires that the ends of the object be marked off simultaneously on our measuring rod as it passes. If we were to mark off the front end and then the back end a little later, the measured length would be too short. It could even be zero; it could even be negative. Clearly, only simultaneous recording of the endpoints will permit a sensible length measurement.

Consider a measurement of the length of the train in the above discussion. The train based observer measures the proper length and can do so at his leisure. That he is moving with respect to the ground is irrelevant. He is in an inertial frame, which is as good as any other. However, the ground based observer must mark the ends simultaneously. If the lightning flashes in the earlier discussion left marks on the tracks then a determination of the distance between those marks would correctly yield the length of the train.

However, an observer on the train is going to say, “Hey, you haven’t recorded the ends of my train simultaneously. You’ve made the front mark before the back one!" Thus he will not recognize the ground based observer's measurement as the length of his train. He will note that in the time between the front mark being made and the back mark being made, the front has moved forward and so the length determined by the ground based observer is going to be shorter than the (proper) length that he measures.

This is the phenomenon of length contraction. Note that it has nothing to do with an actual change in the length of the train as it moves (as was the case with the Lorentz-Fitzgerald hypothesis) but is a simple consequence of the relativity of simultaneity. (The phenomenon is though still often called Lorentz contraction for historical reasons.)

5.4 Relativity is truly relative

Suppose that the train observer wants to measure the length of a section of railway track. Since the track is moving past him, towards the rear of his train, he must mark off the ends of the track simultaneously. But the ground based observer notices that, by his clocks, the front mark is made at a later time than the one at the back, so he is going to say, "Hey, you haven't recorded the ends of track section simultaneously. You've made the back mark before the front one!" Thus he will not recognize the train based observer's measurement as the length of the section of track. In the time between the back mark being made and the front mark being made, the measuring rod in the train has moved forwards and so the length determined by the train observer will be less then the (proper) length measured by the ground based observer.

In both cases (i.e. measurement by train based or ground based observer), the ordinary length of an object, moving with respect to the observer, turns out to be less than the object's proper length.

5.5 The train in the tunnel paradox

Consider a very fast train going through a tunnel. Let both train and tunnel have the same proper length, L_0.

In the tunnel frame the train will appear contracted and so will fit inside the tunnel. However, in the train frame it is the tunnel which is contracted and so the train will not fit. How can this be? Surely the train will either fit or not fit. Something must be wrong.

The resolution of this paradox lies in the meaning of the phrase "fit inside." In saying that the train fits inside the tunnel, a railway worker in the tunnel frame is merely noting that at the same instant of time, both front and back ends of the train are located inside the tunnel. However, his notion of simultaneity differs from that of a passenger on the train. If the railway worker were to turn on a couple of lamps, one at each end of the train, at an instant he considered the train to be "in the tunnel" then the passenger will believe that the rear end of the train has not yet entered the tunnel when the front lamp is turned on. By the time the passenger considers the back lamp to have been turned on, the front of the train will have exited. There is therefore no inconsistency in the passenger's measurement of a short tunnel length and the railway worker's claim that the train fits inside the tunnel. These things are just frame dependent perceptions.

5.6 Transverse lengths

Consider the measurement of the width of the train in the above discussion. Since there is no sideways motion of the train with respect to the ground, a train based observer and a ground based observer will agree on the simultaneity of events occuring directly opposite each other on each side of the train. There is therefore no impact on the measurement of the width of the train.

Neither observer need even mark off the sides of the train simultaneously. One could, for example, arrange for a couple of cans of spray paint to be attached to each side and to squirt paint onto the ground. Two parallel lines will be formed and the ground based observer can measure their separation at his leisure, just as a passenger in the train can do. Both observers measure a proper length for the width of the train.

Similarly, the train passenger can measure the separation of the railway tracks. In fact the separation is clearly that of the wheels on the train. The wheels being attached to the train, their separation can be measured at leisure.

Thus all lengths transverse to the direction of motion are seen to be unaffected by that motion.

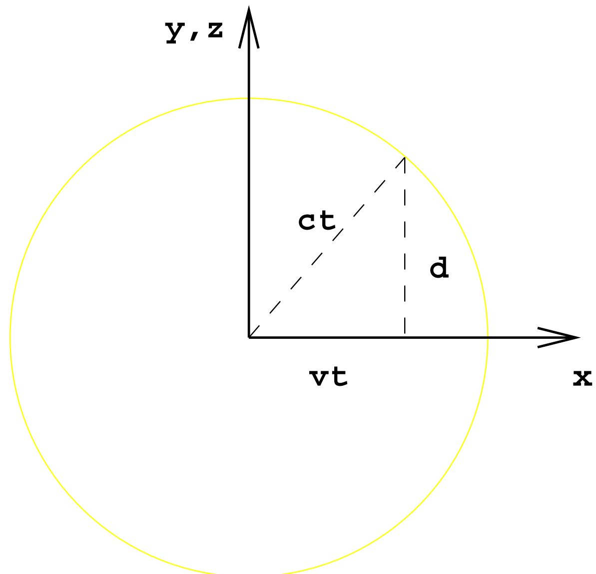

Figure 5.3: A light flash expands as a sphere in every direction and lengths, d, transverse to the direction of motion are unaffected by the relativity of simultaneity.

5.7 Invariance of the spacetime interval

The invariance of transverse lengths has an immediate consequence of much practical importance. Consider a flash of light emanating from a point. Since light travels with the same speed in every direction, then, at any time, the flash of light will have travelled the same distance in every direction, i.e. the flash takes the form of an expanding sphere of radius $c\Delta t$ where Δt is the time from the instant of emission. Consider this projected on an xy-plane and let the source be moving along the x axis with speed v. See fig. 5.3. By Pythagoras' theorem, the transverse distance d satisfies

$$d^2 = (c\Delta t)^2 - (\Delta x)^2. \tag{5.4}$$

But this has to be true for any inertial frame moving along the x direction in the frame of the figure. If primes denote the coordinates in that frame then we must have

$$(c\Delta t')^2 - (\Delta x')^2 = (c\Delta t)^2 - (\Delta x)^2. \tag{5.5}$$

The quantity $(c\Delta t)^2 - (\Delta x)^2$ is often called the *spacetime interval* and we see that it must have the same value in every inertial frame.

5.8 Length contraction

The invariance of the spacetime interval enables us to determine by just how much lengths of moving objects seem contracted in comparison to their proper lengths.

Consider fig. 5.2. The spatial separation of the ends of the moving train is just the proper length in the train frame. If we take the ends at the times t_1 and t_2 then the times are the same in the train frame and the spacetime interval in the train frame is $-L_0^2$. However, in the ground based observer's frame it is (recalling the discussion leading to Eq. 5.1

$$(c\Delta t)^2 - (\Delta x)^2 = c^2(t_2 - t_1)^2 - c^2(t_1 + t_2)^2. \tag{5.6}$$

From Eqs. 5.2 and 5.3 we thus obtain

$$L_0^2 = \left(\frac{c^2}{c^2 - v^2}\right) L^2 \tag{5.7}$$

Hence

$$L = L_0/\gamma \tag{5.8}$$

where

$$\gamma \equiv \frac{1}{\sqrt{1-\beta^2}} \tag{5.9}$$

and

$$\beta \equiv v/c. \tag{5.10}$$

For small velocities, γ is seen to be very close to 1. This is why the effect is not readily apparent. However, as v approaches c the factor γ can become quite large.

Example Consider a stick of proper length L_0 lying in the xy-plane and inclined at an angle θ to the x-axis. The component

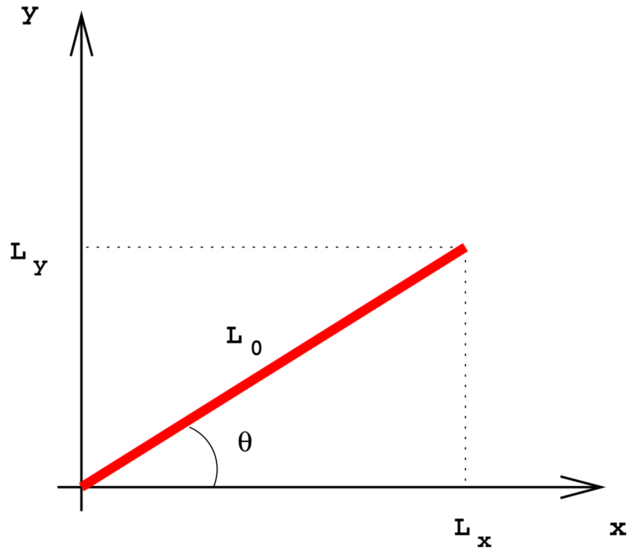

Figure 5.4: Stick at rest.

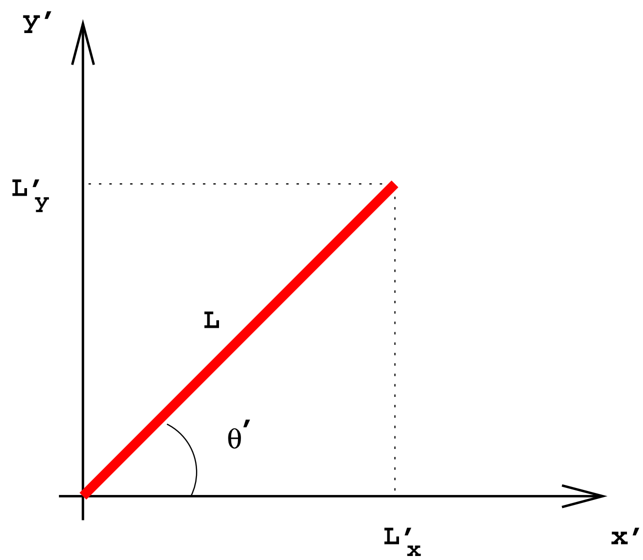

Figure 5.5: Stick moving.

of the stick along the x-axis is thus $L_x = L_0 \cos\theta$ while the component of the stick along the y-axis is $L_y = L_0 \sin\theta$. See the figure. Now consider how this stick appears in a frame in which it is moving with speed v in the x direction. The y component is transverse to the motion and is unaffected:

$$L'_y = L_y.$$

However, the x-component is Lorentz contracted:

$$L'_x = L_x/\gamma.$$

Now consider the angle the stick makes with the x-axis in this frame. It is given by

$$\begin{aligned} \tan\theta' &= \frac{L'_y}{L'_x} = \gamma \frac{L_y}{L_x} \\ &= \gamma \tan\theta. \end{aligned} \tag{5.11}$$

Thus the stick is tilted further away from the x-axis in this frame.

Exercise 5.2 *Show that the length of the moving stick is*

$$L = L_0 \sqrt{1 - \beta^2 \cos^2\theta}.$$

Exercise 5.3 *Show that*

1. *A surface area A remains unchanged in inertial frames moving at constant speed in a direction perpendicular to the surface.*

2. *A surface area A is contracted to A/γ in inertial frames moving at a constant speed in a direction parallel to the surface.*

3. *Any volume V is contracted to V/γ in inertial frames moving at constant speed v in any direction.*

5.9 The visual appearance of rapidly moving objects

It is important to understand that when we talk of Lorentz contraction, etc., we are referring to what is measured, not what we actually see (from afar) with our eyes. To understand the latter we must take into account the finite travel time of light. When we "see" a moving object, what we actually see is the object where it was at some time in the past. (This is analogous to looking overhead at a jet plane and noticing that the sound from it is coming from a point considerably behind. This is because the speed of sound is much slower than that of light.) Furthermore, light from different parts of an object may have to travel (slightly) different distances to our eye and thus take different times. Consequently, objects can appear distorted.

Of course, in practice a very high speed object will appear as just a blur to us. However, these remarks can be interpreted as applying to what would be recorded by a high speed camera.

> **Example** Consider a rod of length L lying in the xy-plane and parallel to the y-axis. In a frame moving to the right along the x-axis with speed v, the rod will appear to be moving to the left with speed $-v$. In this new frame, the length is still L and the measured orientation is again parallel to the y-axis (according to Eq. 5.11). However, what an observer actually sees, depends on the observer's location. Let an observer be located at a height h above the origin. The time taken for light to reach the observer from a point (x, y) in the plane of the rod is
>
> $$\Delta t = R/c$$
>
> where
>
> $$R = \sqrt{x^2 + y^2 + h^2}.$$
>
> Thus at any instant t the observer sees light from those points as they were a time Δt ago. For a point on the rod, h and y do not vary. However, x varies according to
>
> $$x = x_0 - |v|t.$$
>
> If the $y = 0$ end of the rod passes through the origin at time $t_0 = -h/c$ (so that the observer sees it pass at $t = 0$) then
>
> $$x = -|\beta|h - |v|t,$$
>
> for any point on the rod. (Note that if the rod had been tilted, or other than straight, then the time variation of x would have

depended on y.) Thus at any instant t, the observer sees light from those points on the rod with x-coordinate

$$x = -|\beta|h - |v|(t - \Delta t) = -|\beta|(h + ct) + |\beta|\sqrt{x^2 + y^2 + h^2}.$$

Solving, one finds

$$y^2 = (\frac{1-\beta^2}{\beta^2})x^2 + \frac{2x}{\beta}(h + ct) + 2hct + (ct)^2 \text{ for } y < L.$$

At $t = 0$ this reduces to

$$y^2 = (\frac{1-\beta^2}{\beta^2})x^2 + \frac{2x}{\beta}h.$$

For $h = 0$ this implies

$$y = \frac{1}{\gamma\beta}x$$

which represents a somewhat tilted rod. For $h \neq 0$ though, and for general times, the rod appears bent!

Exercise 5.4 *Show that light reaching the observer in the above example, from equally spaced times in the past, comes from points on the xy plane that lie on concentric circles whose radii become evenly spaced a large distance away but are somewhat more widely spaced close to the observer. Sketch these circles for $c\Delta t = h$, $2h$, $3h$, $4h$.*

The locus of points where the moving rod intersects these circles corresponds to what is actually seen. If the rod is moving at $\beta = -0.5$, sketch the position of the rod at $ct = -h$, $-2h$, $-3h$, $-4h$. Thus show what the rod looks like at $t = 0$. What does the rod look like at $t = -h/c$?

Repeat for $\beta = -0.75$.

Review

What does it mean when we say that two events are simultaneous?

In Einstein's train paradox, which end of the train does a person on board think gets struck first?

On a spacetime diagram for a stationary observer, sketch the lines of simultaneity of an observer moving to the right. Sketch another diagram for an observer moving to the left.

What does it mean when we say that something fits inside something else?

What is the length contraction formula? What is the factor γ in that formula?

In which direction does a moving stick appear tilted?

Questions

1. According to legend (likely apocryphal), Galileo dropped two objects, one light and one massive, from the Leaning Tower of Pisa and found that both hit the ground at the same time. This result is considered so important in the theory of gravity that when astronauts visited the Moon they repeated this experiment with a feather and a heavy object. The astronaut performing the experiment reported that, upon being dropped at the same instant from the same height, both fell to the surface at the same rate and hit it simultaneously. If the experiment had been observed by an astronaut zipping past at $0.8c$ in a moon buggy, would the objects still reach the surface of the Moon simultaneously? What would that imply about the equality of the rates of fall for the two objects?

2. Does length contraction imply that moving objects will actually *look* shorter? Can length contraction be visually seen? (Consider what an object looks like as it approaches from afar, as it passes by, and as it recedes into the distance. Consider a close approach and a distant approach.)

3. One method of measuring distance is to roll a wheel along the ground and count the number of times it turns. If the wheel rolls without slipping then the distance the axle moves is C times the number of turns, where C is the circumference of the circle. (Convince yourself that this is true. If you think the issue trivial, how far would a rod placed on top of the wheel advance if it moves without slipping?) This

method is often used by highway departments and is the basis of the odometer in automobiles.

Does the distance measured by this method depend on the speed at which the wheel moves?

The circumference of a circle is related to its diameter, D (or radius, R), by the famous relation,

$$C = \pi D, \ (= 2\pi R),$$

where $\pi = 3.141592654\ldots$ is a constant, known to the Ancient Egyptians as approximately 22/7 but later recognized by the Ancient Greeks as an irrational number. Does the wheel shrink as it spins faster (i.e. does its diameter decrease)? Does the number of turns it makes depend on the rate at which it spins, i.e. would an observer sitting on the axle count the same number of turns as an observer standing on the ground? How does one reconcile the distance measured with the phenomenon of length contraction?

Problems

1. Two gunslingers are having a duel in the center of the main street. As witnessed by an observer from the saloon located midway between the pair, they draw their guns simultaneously and shoot each other dead. Justice has been done says this observer! However, the duel is also witnessed by an observer travelling down the street on a super fast hoverboard at $0.8c$. The hoverboarder is travelling in a straight line from gunslinger A towards gunslinger B.

   (a) Which, if either, of the gunslingers does the hoverboarder think drew first (according to a system of synchronized clocks in the hoverboarder's frame)? If one draws first does he not win the duel and survive?

   (b) Discuss briefly, with the aid of a spacetime diagram, the course of events as seen by the hoverboarder — including any deaths. If you think both die, who dies first?

2. Two "star cruiser" spaceships are engaged in a space battle. They pass each other "head on" at relativistic speeds (and sufficiently close that

the time for light to travel the perpendicular distance between them can be neglected in this problem). The spaceships have identical proper lengths. Spaceship A has a laser blaster in its tail end oriented for a broadside shot (i.e. it fires in a direction perpendicular to the direction of motion). The captain of A gives her gunner an order to fire a shot at the instant that the front of A passes the tail of the other rocket, B. The gunner sees B Lorentz contracted and thinks that his shot will miss. On the other hand the Captain of B sees A Lorentz contracted, thinks his ship will be hit and panics. Who is correct? Does A's shot hit or miss? (It can't do both because that would be a contradiction; B will either bear the damage of a hit or it won't.) Explain your answer fully.

3. John Macho is a guntoting cowboy. He stands beside a train speeding by at $0.6c$. Seated in the train, beside the window and facing forwards, is Arthur Fats, a rotund man of thickness 80 cm (front to back) in the region of his torso. John Macho holds two pistols 80 cm apart against the side of the train, at the height of Arthur's torso, and at the instant that Arthur's back passes, he fires both simultaneously. In John's frame, Arthur appears Lorentz contracted and so John thinks his shots will miss Arthur — he only intends to scare the bejeezzus out of him. A sheriff riding on the train observes the incident and has John arrested. In the sheriff's frame the distance between John's guns appears contracted (to less than the thickness of Arthur's torso).

   (a) What would be the most appropriate charge: murder (or attempted murder if the bullets only cause flesh wounds), or reckless discharge of firearms and endangerment of life (if the bullets do indeed miss)? Justify with a physics discussion of what actually happens.

   (b) How far apart will the crime investigators (on board the train) find the bullet holes in the train carriage walls? Consider both an investigation done while the train is still moving and another performed after the train has come to a gradual stop.

4. Consider a very fast runner holding a pole of proper length L horizontal to the ground. He runs through a barn, also of proper length L, having large doors at both front and back.

In the barn frame the pole will appear contracted and so will fit inside the barn. However, in the runner's frame it is the barn which is contracted and so the pole will not fit. How can this be? Surely the pole will either fit or not fit. Something must be wrong. Resolve the paradox with a clear discussion of the relativity physics involved.

5. Consider a sheet of cardboard from which a pog is cut. Let the cardboard sheet be held horizontally, at rest, and let the pog also be horizontal, but moving upwards and to the right with constant high speed, v. Let the direction of motion be such that the front edge of the pog will just miss touching the right edge of the hole as it moves towards the cardboard sheet. In the frame of the cardboard, the pog appears Lorentz contracted and so should easily pass through the hole. However, in the frame of the pog, the hole is Lorentz contracted and so the pog will not fit. Something must be wrong; the pog will either pass through the hole or it won't. Resolve this paradox.

Chapter 6

Time Dilation

6.1 Time dilation

The invariance of transverse dimensions has important implications for the behavior of clocks. Consider a train pulling a playboy's carriage. The carriage has a mirror on the ceiling and a light pulse is fired from the floor directly upwards in the playboy's frame. It bounces off the mirrored ceiling and returns directly down to the floor a time $\Delta\tau$ later. If the ceiling is of height h, the distance travelled by the light pulse is $2h = c\Delta\tau$. Now consider the trajectory of this light pulse as observed by someone outside of the train, at rest with respect to the ground. Ignoring minute gravitational effects, the light pulse must describe a triangular path. See fig. 6.1. (For those readers bothered by gravity, just construct a similar scenario with rockets in the farthest reaches of outer space.) In the ground based observer's frame, the train moves $\Delta x = v\Delta t$ during the time of the pulse's round trip. Thus the total distance travelled by the light is $c\Delta t > 2h$. But this observer also measures the height of the ceiling to be h and by Pythagoras' theorem

$$h^2 = (c\Delta t/2)^2 - (v\Delta t/2)^2. \tag{6.1}$$

The invariance of transverse lengths thus implies that

$$(c\Delta\tau/2)^2 = (c\Delta t/2)^2 - (v\Delta t/2)^2. \tag{6.2}$$

Thus

$$\Delta t = \gamma\Delta\tau \tag{6.3}$$

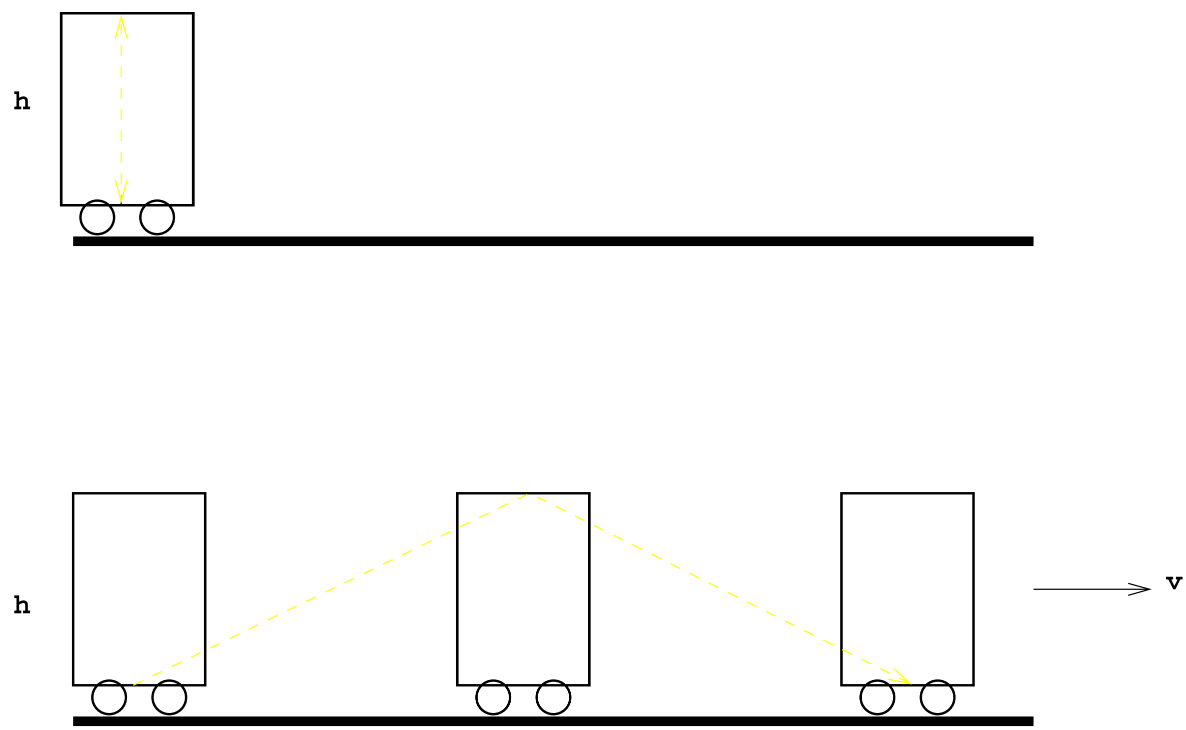

Figure 6.1: A light pulse is bounced off the ceiling of a railway carriage and observed from two different frames. In the top, the train is at rest with respect to the observer; in the bottom, the train is moving with v with respect to the observer. The light pulse is represented by the dashed line.

where γ is the same factor as in the length contraction phenomenon.

Note that the coordinate time Δt measured between two spatially separated events is larger than the proper time $\Delta\tau$. This effect is called time dilation. Hence we see that standard clocks in different frames not only have different synchronizations but also run at different rates.

6.2 Experimental evidence

Nowadays there exists a wealth of evidence in support of the time dilation effect. We note here some of the more important experiments.

6.2.1 Transverse Doppler effect

The first direct observation of time dilation was made by Ives and Stilwell [21] in the 1930s. Their experiment basically looked for a transverse doppler shift (to be discussed further in Chapter 7). If a source is moving towards an observer then the frequency of radiation will be increased whereas if it is moving away from the observer, the frequency will be decreased. These are simple effects related to the lesser or greater distance the light has to travel. However, no such effect is expected at the instant a source is moving past an observer and any frequency shift in those circumstances must represent a fundamental change in the frequency of the source.

The experiment of Ives and Stilwell confirmed the time dilation effect to within about 3%. A modern experiment [22], using lasers and two photon absorption, agreed with the time dilation factor to within 4×10^{-5}.

6.2.2 Cosmic ray muons

A more dramatic example of time dilation can actually be observed in nature with quite modest equipment [23]. Cosmic rays are very high energy particles, mainly protons, from outer space. When they enter Earth's atmosphere they interact with the nuclei of the atmospheric gas molecules. One of the products of these reactions is a subatomic particle called a muon. In the laboratory, muons decay with a mean life of $\tau = 2.2\,\mu$s. Now suppose that the muon is generated at an altitude of 10,000 m with a speed $v = 0.99c$. This speed corresponds to a gamma factor of 7.1. Were it not for the time dilation factor, the muons would travel, on average, about $v\tau \approx 650\,$m before decaying and thus very few would reach Earth's surface. However, copious numbers of muons can be detected at sea level. With the time dilation factor included, they travel

$$x = vt = v\gamma\tau \approx 4800\,\mathrm{m} \tag{6.4}$$

thus accounting for the numbers reaching sea level.

6.2.3 Particle accelerator experiments

A similar effect occurs in high-energy particle accelerator laboratories. Since many subatomic particles are unstable and live for only very short times, it is not possible to accelerate them. Instead, one takes stable particles, such as the proton or electron and accelerates those. This primary beam is then aimed at a target. Collisions with the target generate a high speed secondary beam of unstable particles. This secondary beam travels downstream to the experimental area. Were it not for the time dilation effect, this beam would typically travel only a few cm before decaying and the luminosity of the secondary beam at the experimental area would be too low to perform experiments.

In one particularly interesting experiment, Bailey et al. [24] measured the half-life of muons circulating in a storage ring. The muons were circulating at $v = 0.99942c$ and the half-life was found to be 29.3 times greater than that of muons at rest. The rest frame of the muons in this experiment is of course not inertial (a point to which we shall return) but nevertheless this corresponds well (to within 1×10^{-3}) with the γ factor for this high velocity.

6.2.4 Thermal motion

Thermal vibrations can cause atoms to move quite fast, of the order of 1 km/s. While the γ factor for such velocities is still quite small, thermal effects have been detected by Pound and Rebka using the Mossbauer effect.

The Mossbauer effect involves essentially recoilless emissions from atoms in solids. (The atom recoils against the entire solid and the huge relative mass of the solid renders recoil entirely negligible.) A useful feature of Mossbauer emissions is that the spectral lines have a very well defined frequency with a very narrow width. Normally, because of recoil, the emission line is at a slightly different frequency than that necessary for absorption (due to energy loss in recoil). However, with recoilless emission, the absorption frequency is naturally matched; any slight change in the emission frequency will impede absorption.

Pound and Rebka used gamma rays from Fe nuclei to study the emission frequency. Heating a sample, thus making the atoms vibrate faster, causes a lengthening of the period, which is a measure of time. Since frequency is the reciprocal of the period, a decrease in the frequency should result. This was indeed observed.

6.2.5 Flying atomic clocks

We have previously noted the very important experiment of Hafele and Keating [10] in which a very accurate atomic clock was flown in an airplane and compared with one left on the ground. The flying clock was found to run slower because of the velocity effect. (Most of the observed effect comes from the difference in gravitational potential between the clocks but this can be calculated and both effects are found to be necessary to obtain agreement with the data.)

6.3 Connection with length contraction

If one considers the cosmic ray muon experiment one could also analyze it from the point of view of Lorentz contraction in the muon's frame. In this frame, Earth is rushing up to meet the muon and the muon's observation of its altitude, h, above Earth is subject to length contraction. The measured velocity of Earth is thus

$$v_E = h/\tau = (h_0/\gamma)/\tau = h_0/(\gamma\tau) = h_0/t \tag{6.5}$$

But this is just the velocity of the muon as measured in the Earth frame.

This therefore establishes the equality of the two methods of measuring distance discussed in chapter 2.

6.4 Travel to the future

The human heartbeat can be considered as a crude clock. Since the heart resides inside a person, it measures proper time for that person. Similarly, all other body cycles are crude measures of proper time for their owner. Included is the aging process. (An average person lives for about 70 years and thus this can be considered as a very crude unit of time. In fact a similar unit — from birth to procreation — is sometimes used and thus primitive peoples often speak of an important event happening a certain number of generations ago.)

Suppose then that a person goes on a high speed journey. The time as measured by someone staying at home will be dilated in comparison with the proper time of the traveller. Therefore the traveller will age less than the person staying at home. If the journey is fast enough and long enough, the traveller can in principle travel to the distant future.

Unfortunately, perhaps, the speed required to achieve this feat in any significant way is enormous. The highest speeds relative to Earth achieved to date by man occurred for the astronauts who went to the Moon. In that case $\beta \approx 4 \times 10^{-5}$. For a jet pilot, β is only about 10^{-6}. As can be seen from the table, one requires a value of β very close to 1 to get a marked time dilation effect.

β	γ
0.01	1.00005
0.1	1.005
0.5	1.15
0.6	1.25 (=5/4)
0.8	1.667 (=5/3)
0.9	3.3
0.99	7.1
0.999	22.4

Example [Trip to Andromeda] The Andromeda galaxy is about 2,000,000 ly away from Earth. How fast would an astronaut need to go to get there while aging only 20 years?

The astronaut will have to travel very close to the speed of light. Thus in Earth's frame the trip will take about 2,000,000 years. One seeks

$$\gamma = \frac{2,000,000}{20} = 10^5$$

in order to achieve the desired time dilation effect. Since

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

this implies that

$$\beta = (1 - 10^{-10})^{1/2} = 0.99999999995.$$

(For those who don't have super accurate calculators, or have simply left them elsewhere, it is very useful to apply the binomial theorem which states that

$$(1 - x)^{1/2} \approx 1 - \frac{1}{2}x + \cdots \tag{6.6}$$

For x very small, higher order terms in the (Taylor) series are negligible.)

6.5 The twin paradox

The possibility of time travel to the future presents the following (in)famous paradox.

Consider two identical twins, Tom and Jerry. Jerry takes a high speed space trip while Tom stays at home. In Tom's frame, Jerry travels at $\beta = 0.8$ for 12.5 years and then quickly turns around and returns home at the same speed, arriving back 25 years after departure. But the proper time experienced by Jerry is

$$\tau_{Jerry} = t/\gamma = 25/(5/3) = 15\,\mathrm{y}.$$

Thus Jerry will appear 10 years younger than Tom, though they were both born at the same time.

However, Tom's age is also a proper time and so Jerry, who sees Tom on Earth receding away and then returning, thinks that Tom has aged

$$\tau_{Tom} = 15/\gamma = 9\,\mathrm{y}.$$

But Tom can't be both 6 years younger and 10 years older than Jerry. Something is wrong.

The problem here is that while Tom is in an inertial frame for the entire episode, Jerry is not. He must decelerate in order to turn around and during that period his inertial frame detector would perceive the deceleration. Thus his calculation of Tom's age is invalid. Tom's account is correct: Tom ages 25 years while Jerry ages just 15.

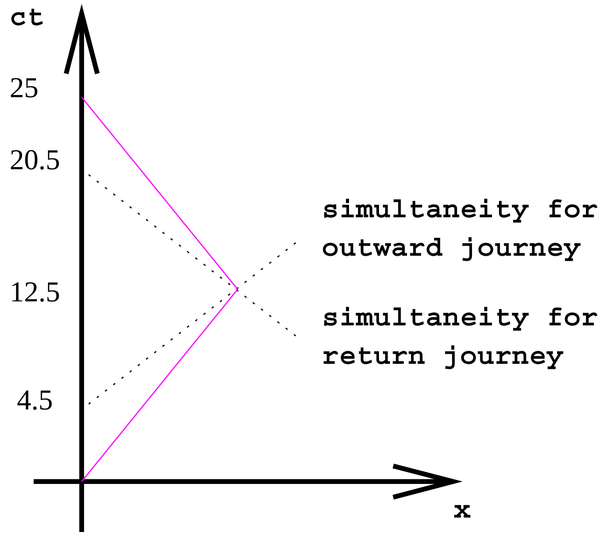

Figure 6.2: Spacetime diagram for the twin paradox.

The spacetime diagram in fig. 6.2 aids in understanding the situation. During the outward journey, Jerry is indeed in an inertial frame and there is nothing wrong with his assertion that Tom ages only 4.5 years. Similarly, Jerry is in an inertial frame on the return trip and correctly notes that Tom ages another 4.5 years. However, at the point where he turns around, the lines of simultaneity rapidly change and thus the Earth clocks seem to make a jump (or speed up tremendously).

This analysis assumes that nothing drastic occurs to Jerry's clock as he turns around. For a huge deceleration the clock may well be destroyed. (Jerry certainly can't withstand accelerations more than a few times that due to gravity at Earth's surface.) However, in principle a clock certainly does keep running during periods of acceleration and its rate of timekeeping is not drastically altered. To account for such situations we shall adopt the *clock hypothesis* which states that an accelerating clock keeps the same time as a clock in an inertial frame with the same instantaneous speed. (The storage ring experiment of Bailey et al. [24] provides direct evidence for this.) This same result could be obtained in the above problem by simply asserting that in fact Jerry doesn't return but instead another person returns in his place in his own spaceship. Thus all travellers can be assumed to be in inertial frames. As they pass each other, Jerry simply signals his time to the other, who sets his clocks accordingly.

6.6 Metric

One notes in all of this that the curved path on a spacetime diagram experiences a smaller elapsed time than the vertical path of a stationary person. This is in contrast to the distance travelled, which is always least for a straight line. However, here we are basically dealing with the spacetime interval

$$c\tau = \sqrt{(ct)^2 - x^2} \tag{6.7}$$

and the minus sign means that the measured proper time is less along a curved path:

$$\tau = \int d\tau = \int \frac{\sqrt{(cdt)^2 - dx^2}}{c} = \int \sqrt{1-\beta^2}\, dt = \int \frac{dt}{\gamma} < \int dt. \tag{6.8}$$

This thus differs from the distance measurement (which in two dimensions is):

$$s = \int ds = \int \sqrt{dx^2 + dy^2} = \int \sqrt{1 + (dy/dx)^2}\, dx > \int dx, \tag{6.9}$$

and involves a positive sign (in the distance interval).

We acknowledge this by saying that the spacetime of inertial frames has a *metric* with *signature* $(+,-,-,-)$ as opposed to the $(+,+,+)$ for ordinary space.

Review

Which time is dilated: ordinary (coordinate) time or proper time?

What is the equation relating coordinate time to proper time?

What direct experimental evidence exists for time dilation?

In the twin paradox, which twin ages less: the traveller or the stay-at-home twin?

In the twin paradox, why are the viewpoints of the twins not symmetric?

Questions

1. Light from the Andromeda galaxy reaches Earth approximately two million years after it was emitted by a star in Andromeda. In the frame of a photon from Andromeda, how long does the journey take? In the photon's frame, how far was the distance travelled?

2. In the last chapter we noted that length contraction did not mean that the length of a moving object really changed but simply that one's measurement of length was affected by the relativity of simultaneity. Is time dilation therefore "real"?

Problems

1. In the movie *Star Wars* fabulous futuristic spaceships move about the universe by making a "jump to light speed". Suppose such a spaceship makes the journey to a star system 150 light years away by this means (i.e. by travelling at the speed of light). As measured by the travellers on board this spaceship, how long will it take the spaceship to arrive at its destination? By how much will the occupants age during the trip? Will they have time to play a game of chess?

2. The normal gestation period for humans is 39 weeks. A space colonist conceives on the eve of her departure from Earth. Her spacecraft accelerates very quickly and then travels at a constant speed of $0.8c$ towards its destination, 10 light-years away. Assume a normal pregnancy (and let 1 year be 52 weeks exactly).

   (a) How old will the space colonist be (in years and weeks) when she gives birth if she departs on her 20th birthday?

   (b) If the mother-in-law insists on watching the birth live on TV — yes, some people have no class — how long after departure must she wait to see the birth?

   (c) How old will the child be on arrival?

3. The theory of relativity often inspires the phrase "moving clocks run slow." Discuss the validity of this statement.

4. A bullet train is moving along some tracks at a speed v which is a significant fraction that of light. A stuntman is riding a motorbike along the roof of the train at a speed $-v$ in the opposite direction so that he is at rest relative to the ground. (This statement is correct in both the frame of a ground based observer and a passenger on the train. See chapter 10 for formal verification.) A ground based observer notes that clocks on the train are running slower than his. The train based observer in turn notes that the stunt rider's clocks are running slow relative to his. But the stunt rider and the ground observer are in the same frame and so their clocks should run at the same rate. How can this be if the stunt rider's clocks are slower than the train observer's clocks, which are slower than the ground-based observer's?

Chapter 7

The Doppler Effect

The doppler effect is a well-known phenomena involving the frequency of radiation received from a source in motion relative to the observer. In the case of sound waves it is manifested by the sound of an approaching vehicle being of higher pitch than a receding one. A very similar effect occurs for light. However, it is simpler in its detail because sound travels in a medium and the velocity of source and observer relative to the medium is important whereas with light only the relative velocity of source and observer matters.

7.1 The doppler frequency shift

7.1.1 Relative approach of source and observer

Consider an observer (at the origin for simplicity) and a source approaching that observer. The source emits a flash of light every τ seconds, as measured in the source frame. This is a proper time and so the period of the flashing in the observer's frame is $t = \gamma\tau$.

The dilated period t is not though the time between flashes as *seen* by the observer. This time is less because, after emitting the first flash, the source moves towards the observer and so is closer when the second one is emitted. This means that the second flash doesn't have as far to travel and thus gets there quicker. See fig. 7.1.

If the source is moving with velocity v then the distance travelled between emission of consecutive flashes is

$$\Delta x = vt = \beta ct. \tag{7.1}$$

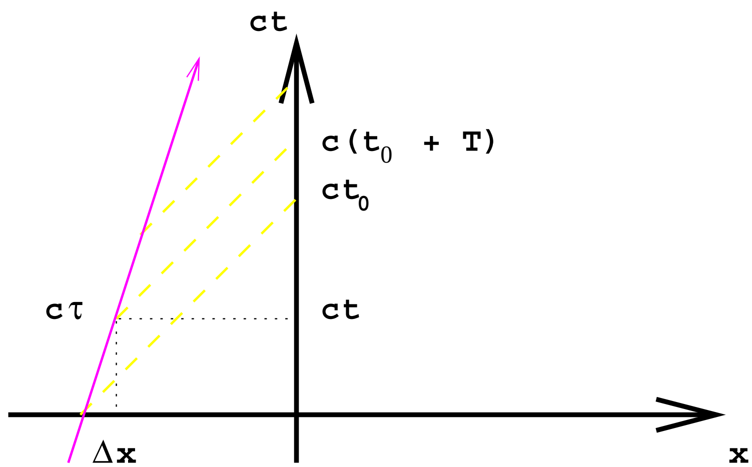

Figure 7.1: A source emits pulses of light a time τ apart in the source frame. The pulses are received in the observer's frame a time T apart, which is less than the dilated period because the next pulse is emitted closer when the source is approaching.

Thus the time between received flashes is

$$\begin{aligned} T &= t - \Delta x/c \\ &= t - \beta t \\ &= \gamma(1-\beta)\tau. \end{aligned} \tag{7.2}$$

The frequency of received pulses is just the reciprocal of the period. Thus in the source frame it is $f_s = 1/\tau$. In the observer's frame the frequency is

$$f = \frac{1}{T} = \frac{f_s}{\gamma(1-\beta)} = \sqrt{\frac{1+\beta}{1-\beta}} f_s. \tag{7.3}$$

This result is immediately applicable to the regular oscillations of light. It means that the frequency of light as seen by an observer is increased from its value in the source frame when the source and observer are approaching each other. Since blue light is of higher frequency than other parts of the visible spectrum we say that the light is *blue-shifted*.

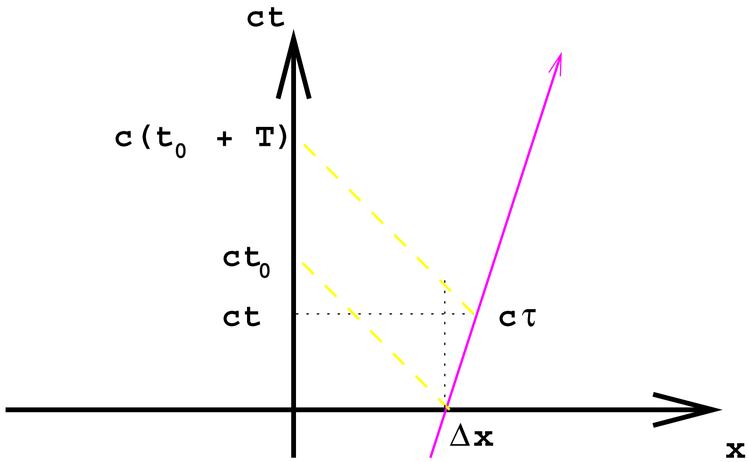

Figure 7.2: A source emits pulses of light a time τ apart in the source frame. The pulses are received in the observer's frame a time T apart, which is greater than the dilated period because the next pulse is emitted further away when the source is receding.

7.1.2 Source receding from observer

The analysis in the case where the source is receding from the observer is analogous. The periods between flashes are again related by the time dilation expression $t = \gamma\tau$. However, because the source is now receding, the second of two consecutive flashes is now emitted further away and thus the light has further to travel. See fig. 7.2. Accordingly the time between received flashes is

$$\begin{aligned} T &= t + \Delta x/c \\ &= t + \beta t \\ &= \gamma(1+\beta)\tau. \end{aligned} \tag{7.4}$$

and the frequency is

$$f = \frac{1}{T} = \frac{f_s}{\gamma(1+\beta)} = \sqrt{\frac{1-\beta}{1+\beta}} f_s. \tag{7.5}$$

where again $f_s = 1/\tau$ is the frequency in the source frame.

This time the observed frequency is less than the emitted frequency and light is said to be *red-shifted*.

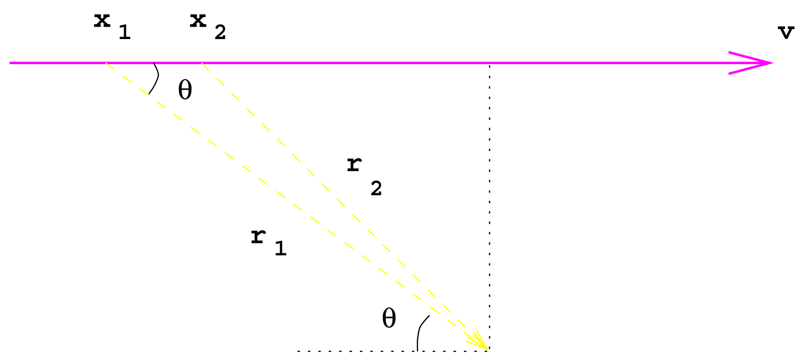

Figure 7.3: A source approaches an observer but not in head-on fashion.

7.2 Transverse doppler effect

If the source is not moving directly towards or away from the observer then the above results need generalizing. Consider fig. 7.3.

Let light pulses be emitted at t_1 and t_2 at x_1 and x_2 respectively. The time interval between emission of pulses is thus

$$t_2 - t_1 = t = \gamma\tau. \tag{7.6}$$

Since the pulses have to travel distances r_1 and r_2 to reach the observer, the time difference on arrival is

$$T = (t_2 + r_2/c) - (t_1 + r_1/c) = t - (r_1 - r_2)/c. \tag{7.7}$$

In most practical situations, the distance from observer to source is much greater than $\Delta x \equiv x_2 - x_1$. Thus

$$r_1 - r_2 \approx \Delta x \cos\theta \tag{7.8}$$

and since $\Delta x = vt$, where v is the source speed in the observer's frame, we have

$$\begin{aligned} T &= t - \Delta x \cos\theta/c \\ &= t(1 - \beta\cos\theta) \\ &= \gamma(1 - \beta\cos\theta)\tau. \end{aligned} \tag{7.9}$$

As before, the frequency is just the reciprocal of this:

$$f = \frac{f_s}{\gamma(1 - \beta\cos\theta)}. \tag{7.10}$$

For $\theta = 0°$ we recover the earlier results for a source approaching directly and for $\theta = 180°$ we recover the results for a source receding.

The case $\theta = 90°$ is interesting. Then, the effect is due solely to the time dilation factor. Unfortunately an experiment to test this is difficult to conduct as a slight misalignment will greatly affect the result.

In 1938, Ives and Stilwell used a clever method to circumvent the difficulties in measuring the doppler effect at exactly 90°. Instead they measured the effects at *both* 0° and 180°. The average observed frequency is

$$\frac{f_0 + f_{180}}{2} = \frac{\gamma(1+\beta)f_s + \gamma(1-\beta)f_s}{2} = \gamma f_s. \tag{7.11}$$

The experiments of Ives and Stilwell agreed with the γ factor to within 3% and played an important historical role in confirming the validity of the time dilation effect.

Note that if we expand γ for small β:

$$\gamma = (1 - \beta^2)^{-1/2} \approx 1 + \frac{1}{2}\beta^2 + \cdots \tag{7.12}$$

we see that the frequency shift varies quadratically with β in leading order, while for the ordinary doppler effect there are terms proportional to β:

$$[\gamma(1-\beta)]^{-1} \approx 1 + \beta + \cdots \tag{7.13}$$

Thus the transverse doppler effect — or the related Ives-Stilwell experiment — is often called the "second-order" doppler effect.

7.3 Astronomical doppler effects

If one observes the spectra of light from stars, and measures the frequency of certain readily identifiable spectral lines, then one finds that the frequencies are shifted. The most obvious interpretation is that one is observing doppler shifts due to the relative motion of the stars and Earth.

There are several means of estimating the distances to stars and one finds that apart from a few close ones, all stars have a redshift. This is usually interpreted as implying that the universe is expanding. Furthermore, the further away a star the greater its redshift, an effect that can be explained in terms of uniform expansion.

Thus astronomers find it useful to define a quantity called the *redshift*, by

$$z = \frac{f_s - f}{f} \tag{7.14}$$

which is greater than 1 for receding sources. In the case of receding sources one readily finds

$$\beta = \frac{(z+1)^2 - 1}{(z+1)^2 + 1}. \tag{7.15}$$

Values of β which are a substantial fraction of c are not uncommon.

7.4 Doppler broadening of spectral lines

Atoms emit light of characteristic frequencies. When dispersed by a prism (or more sophisticated spectroscope) these characteristic frequencies manifest themselves as narrow bands of light in discrete colors and are therefore known as spectral lines. The natural width of these lines, i.e. the frequency range within each line, is fairly narrow. However, if the atom is in a gas at a high temperature it will, on average, possess a high velocity. Because the motion of gas molecules is random, the direction of motion of an individual atom, at the moment it emits light, will sometimes be towards the observer and sometimes away. When the atom is moving toward the observer its spectra will be blue-shifted and when it is moving away its spectra will be red-shifted. Since the observed spectra comes from a multitude of atoms, the observed effect is a broadening of the individual spectral lines.

Accurate measurement of this broadening for ordinary gases is very difficult but it has been measured by Pound and Rebka using the Mossbauer effect for the gamma rays emitted by Fe nuclei.

7.5 Twin paradox

The twin paradox may be conveniently analyzed using the doppler effect. One observes that if each twin emits f_s pulses per unit time then the other receives them at the rate f according to the doppler effect. However, there is an asymmetry because when the traveller turns around he immediately starts seeing pulses from the stay-at-home twin at a different rate. On the other hand, the stay-at-home twin has to wait an extra time L/c after the turning time (where L is the distance to the turning point) before the pulses start arriving at the new rate. If the pulses are added up then each agrees on the other's age and there is no paradox.

Exercise 7.1 *Sketch a spacetime diagram for the stay-at-home twin showing the trajectories of both twins and those of a series of light pulses emitted by each in the direction of the other.*

Review

What is the doppler effect?

Why is the time between consecutive pulses seen by an observer in motion with respect to the source not just the dilated period?

If a light source is moving towards an observer, will the light be blue shifted or redshifted?

What is a second-order doppler effect?

What name is given to the quantity defined as the fractional change in frequency, $\Delta f/f$?

Questions

1. What are the problems associated with trying to measure the transverse doppler shift at exactly 90°?

2. If the spectral lines of atoms in distant stars are frequency shifted, how does one identify them?

3. Does the fact that the color of observed light depends on the relative motion of source and observer imply that there is a preferred frame of reference?

Problems

1. How fast would a person need to drive in a motor car in order for a red traffic light ($f_s = 4.6 \times 10^{14}$) to appear green ($f = 5.7 \times 10^{14}$)?

2. Analyze the twin paradox using the doppler effect. Show that if the twins count the number of light pulses received from the other during the traveller's journey then both twins will agree on the age of the other.

3. Show that a change in the velocity of recession, or approach, of a source leads to a change in the doppler shifted frequency given by

$$\frac{df}{f} = -\frac{1}{\beta}\frac{d\gamma}{\gamma}, \tag{7.16}$$

   where β is positive for a source receding.

Chapter 8

Coordinate Transformations

It is frequently useful to be able to transform the coordinates (x^0, x^1, x^2, x^3) of an event as determined in one inertial frame to the coordinates (x'^0, x'^1, x'^2, x'^3) of another frame. There is nothing magical about this transformation; it is just a mapping from the set of coordinates one observer uses to that used by another observer.

8.1 Translations

The simplest transformation is that for a shift in the origin of the coordinate system. If, for example, we move all of our measuring rods a meters in the positive x direction then all x coordinates will change to

$$x' = x - a. \tag{8.1}$$

Similarly, a shift in the instant chosen as the zero of time will result in a new set of time coordinates,

$$ct \to ct' = ct - a^0. \tag{8.2}$$

Such a shift takes place when one resets a wristwatch back to normal time after the end of daylight savings time, or when one wishes to compare the (Christian) Gregorian calendar with the Moslem one. In general, if a^μ is the shift in origin for a particular coordinate ($\mu = 0, 1, 2, 3$) then the coordinates of all objects will change to

$$x'^\mu = x^\mu - a^\mu. \tag{8.3}$$

This is referred to as a *passive* change of coordinates and is to be distinguished from taking an object and displacing that. The latter type of displacement is called an *active* transformation. It would be described by $x^\mu \to x'^\mu = x^\mu + a^\mu$ and is seen to be the inverse of the passive transformation.

8.2 Rotations

Consider a set of coordinates in an xy plane. They form a graph-paper-like square grid. Now rotate this coordinate system about the (perpendicular) z-axis. Since lengths don't depend on orientation, the scale does not change and the grid still appears as composed of parallel, equally spaced straight lines. Lines of constant x' and y' in the new coordinate system appear in the old one as in fig. 8.1. The transformation of coordinates can be readily

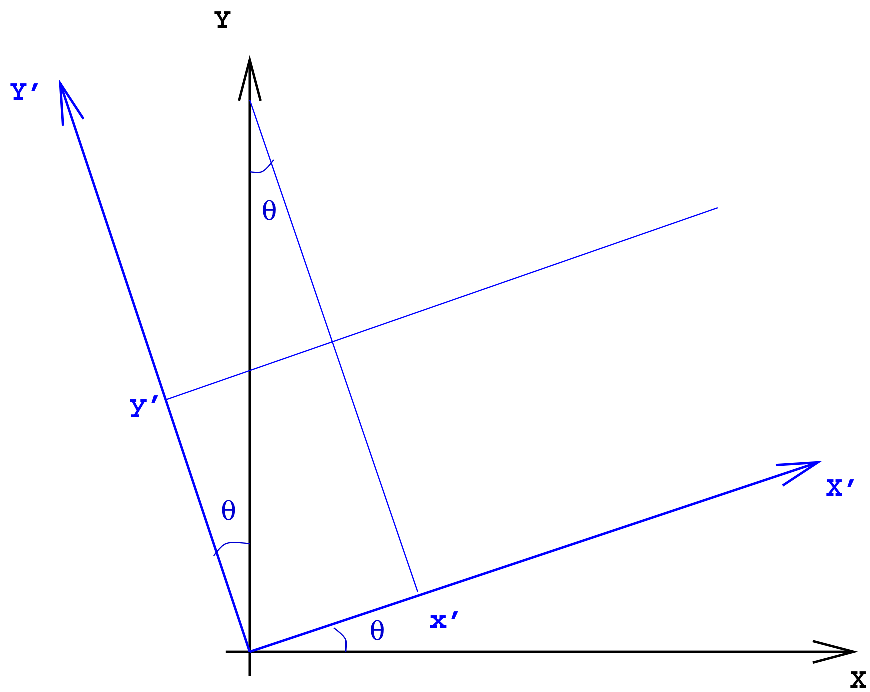

Figure 8.1: The coordinate system is rotated by an angle θ about an axis perpendicular to the XY plane.

obtained by considering the equations for these particular straight lines in the original coordinate system. The equation of a straight line always takes the form

$$y = mx + C \tag{8.4}$$

where m is the slope and C the intercept on the y-axis. Consideration of fig. 8.1 shows that the line of constant y' is given by

$$y = \tan\theta\, x + y'/\cos\theta \tag{8.5}$$

while that of constant x' is given by

$$y = -\cot\theta\, x + x'/\sin\theta. \tag{8.6}$$

Inverting these, one immediately obtains

$$\begin{aligned} x' &= \cos\theta\, x + \sin\theta\, y \\ y' &= -\sin\theta\, x + \cos\theta\, y \end{aligned} \tag{8.7}$$

and of course we also have $z' = z$.

The corresponding active rotation is again just the inverse and is obtained by changing θ to $-\theta$. Since $\cos(-\theta) = \cos\theta$ and $\sin(-\theta) = -\sin\theta$, this results in the sine terms in the above expression changing sign.

Exercise 8.1 *Show that a passive rotation by θ about the x-axis is given by substituting y for x and z for y in Eq. 8.7. Similarly show that a rotation about the y-axis is given by substituting x for y and z for x. (Assume the coordinate system to be the standard right-handed one.)*

In general a rotation can be described by a rotation angle and a rotation axis. The specification of these is equivalent to specifying three rotation angles for an arbitrary rotation. There is more than one way of defining these three angles. We need not bother here with details though.

It is noteworthy that a rotation leaves lengths invariant. Indeed, a rotation can be considered as (partly) defined by this fact.

Exercise 8.2 *Verify that*

$$x'^2 + y'^2 + z'^2 = x^2 + y^2 + z^2 \tag{8.8}$$

under the above transformation.

8.3 The Lorentz boost transformation

Consider a measuring rod moving along the positive x axis with speed v. The worldlines of the graduation marks on the rod trace out equally spaced parallel lines of slope $1/\beta$ on a spacetime diagram. Comparing these worldlines with those of a rod at rest (which move straight up) we see that if the origin is appropriately chosen in the frame of the measuring rod then the $x' = 0$ worldline will be the ct' axis. Similarly, we deduce that if the zero of time is appropriately chosen, then the line of simultaneity, for the rod, passing through the origin, is the x' axis. The homogeneity of space and time means that the equally spaced coordinate grid in the rod frame will appear as an equally spaced, but no longer rectangular, grid of parallel straight lines in the original frame. See fig. 8.2. As with rotations, we seek the equations describing those straight lines.

The line of constant x' has slope $1/\beta$. Its intercept on the x axis is the length of a rod of proper length x' in the rod frame. Thus this intercept must be x'/γ from the length contraction formula. However we need the intercept $-\alpha$ on the y axis. But we know the slope is $1/\beta$ so we must have

$$\alpha = 1/(\beta\gamma). \tag{8.9}$$

Thus the equation of this line is

$$ct = \frac{1}{\beta}x - \frac{1}{\beta\gamma}x'. \tag{8.10}$$

Similarly, the slope of the line of constant ct' has to be β since it is symmetrically located about the path of a light ray with respect to the lines of constant x'. This has to be so because the worldline of the light ray is an identical set of points in both frames and in both frames it must have slope of 1 (corresponding to the velocity of light being c in both frames). Now consider the intercept ct on the y axis. If we consider a clock at the origin in the original frame then this intercept will be the proper time registered by that clock. However, in the rod frame this clock is a moving object which is at a different position at t' than at time 0. Thus $t' = \gamma t$ and the intercept is ct'/γ. We therefore find that the equation of the line is

$$ct = \beta x + \frac{1}{\gamma}ct'. \tag{8.11}$$

Inverting these we obtain

$$\begin{aligned} ct' &= \gamma(ct - \beta x) && (8.12) \\ x' &= \gamma(x - \beta ct) && (8.13) \end{aligned}$$

and of course, from the invariance of transverse lengths we also have

$$\begin{aligned} y' &= y && (8.14) \\ z' &= z. && (8.15) \end{aligned}$$

Eqs. 8.12-8.15 are known as the Lorentz transformations. They describe a passive Lorentz boost from one inertial frame to another moving uniformly with respect to it. (It is to be noted that the transformation is symmetric under interchange of x and ct.)

The active boost again takes the same form as the inverse transformation from the rod frame to the original frame. Since the original frame is moving with velocity $-\mathbf{v}$ in the rod frame, this inverse transformation is simply obtained by changing the sign of β in the above.

Exercise 8.3 *By considering the coordinate system of the original frame on a spacetime diagram for the rod frame, verify that the inverse transformation is given by*

$$\begin{aligned} ct &= \gamma(ct' + \beta x') && (8.16) \\ x &= \gamma(x' + \beta ct') && (8.17) \\ y &= y' && (8.18) \\ z &= z' && (8.19) \end{aligned}$$

and that the active transformation to the rod frame has this same form.

It is to be noted that the Lorentz transformations mix up space and time coordinates. This is why we refer to space and time as a single four-dimensional entity called spacetime.

Example Consider a square at rest in the xy plane with sides of length 1. See fig. 8.3. Make a passive boost along the x direction with $\gamma = 5/4$ (i.e. $\beta = 3/5$). The Lorentz transformations for this read

$$\begin{aligned} ct' &= \frac{5}{4}ct - \frac{3}{4}x \\ x' &= \frac{5}{4}x - \frac{3}{4}ct \\ y' &= y \end{aligned}$$

and thus the (x, y) coordinates change according to

$$\begin{aligned} (0,0) &\rightarrow (0,0) \\ (0,1) &\rightarrow (0,1) \\ (1,0) &\rightarrow (\frac{5}{4},0) \\ (1,1) &\rightarrow (\frac{5}{4},1). \end{aligned}$$

Now the original square was at rest, so its sides were of proper length 1. How can the x coordinate of the right hand ends now be greater than 1 when the square should be Lorentz contracted? The answer lies in the time components. The first two of the above new coordinate pairs (for the left hand end) are at $ct' = 0$ but the second two new pairs are at $ct' = -3/4$. Thus the x coordinates do not correspond to the same time and their difference is not a valid length measurement. In fact we can see that all is consistent with the expected length contraction by working out the x' coordinates of the right hand end at $t' = 0$. In the new frame, the box is moving with $\beta = -3/5$, i.e. in the negative x' direction. It travels $\Delta x' = v\Delta t' = \beta c\Delta t'$ in time t' and so for $c\Delta t' = +3/4$ we have $\Delta x' = -9/20$. Thus the x' coordinate of the right hand end at $t' = 0$ is

$$x' + \Delta x' = \frac{5}{4} - \frac{9}{20} = 4/5$$

and this corresponds correctly to the length contraction factor of $1/\gamma$.

The transformations for boosts along either the y or z axes are easily inferred from the above and are given by substituting y or z for x. A boost in a general direction can be parameterized with three parameters which can be taken as the components of a vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. Using this, an arbitrary Lorentz transformation can be written as

$$x'^0 = \gamma(x^0 - \boldsymbol{\beta} \cdot \mathbf{x}) \tag{8.20}$$

$$\mathbf{x}' = \mathbf{x} + \frac{(\gamma - 1)}{\beta^2}(\mathbf{x} \cdot \boldsymbol{\beta})\boldsymbol{\beta} - \gamma x^0 \boldsymbol{\beta} \tag{8.21}$$

Exercise 8.4 *Show this by noting that if the boost is in the direction*

$$\mathbf{x}_{\|} = (\mathbf{x} \cdot \boldsymbol{\beta})\, \boldsymbol{\beta}/\beta^2 \tag{8.22}$$

then the spatial vector perpendicular to the motion is

$$\mathbf{x}_{\perp} = \mathbf{x} - \mathbf{x}_{\|}. \tag{8.23}$$

Another way of describing a boost in an arbitrary direction is simply to rotate the coordinate system until the x axis is aligned with the boost direction, make the boost, and then rotate back.

The Lorentz boost transformations leave invariant the spacetime interval, i.e.

$$(ct')^2 - \mathbf{x}'^2 = (ct)^2 - \mathbf{x}^2. \tag{8.24}$$

Exercise 8.5 *Verify this.*

This is true for any ct or $\mathbf{x}$. Previously we only had established this invariance for the case where the interval was positive. We see now that it applies to any spacetime interval in the frame of an inertial observer who is travelling at a speed less than c. Note also that since rotations leave lengths invariant they also leave the spacetime interval invariant.

8.3.1 Low velocity (Galilean) limit

For $v \ll c$, but not necessarily zero, we note that $\gamma \approx 1$. Thus the Lorentz transformations reduce to

$$\begin{aligned} t' &= t \\ x' &= x - vt \\ y' &= y \\ z' &= z. \end{aligned} \tag{8.25}$$

Exercise 8.6 *Show this, assuming that x is relatively small in comparison with ct/β.*

The equality of times in the two frames in this limit is responsible for the common perception, espoused by Newton, that time is universal. We see now though that this is only an illusion created by the high value of c.

Eqs. 8.25 are frequently referred to as the Galilean transformations. It is worth emphasizing that for very large space separations, the Galilean transformations must be corrected even for $v \ll c$. The relativity of time is a truly inescapable consequence of a finite value of c and its constancy in all inertial frames.

8.3.2 Rapidity

Two boosts in the same direction are conveniently described using the *rapidity*, η, defined using hyperbolic functions as

$$\begin{aligned} \cosh\eta &= \gamma \\ \sinh\eta &= \beta\gamma. \end{aligned} \tag{8.26}$$

Exercise 8.7 *Verify that this definition of η satisfies the hyperbolic identity*

$$\cosh^2\eta - \sinh^2\eta = 1. \tag{8.27}$$

When one makes a boost by η_1, followed by a boost by η_2 in the *same direction*, this is equivalent to a single boost by

$$\eta = \eta_1 + \eta_2. \tag{8.28}$$

This is analogous to making a rotation by θ_1 followed by a rotation by θ_2 about the same axis; the result is equivalent to a single rotation by $\theta = \theta_1 + \theta_2$ about that axis.

8.3.3 Wigner rotation

Though mathematically messy, two rotations about any axis are always equivalent to a single rotation about some axis. However, the situation with boosts is more complicated. In general, two boosts are not equivalent to a single boost but rather must be described as a combination of both a boost and a rotation.

Example Consider the square in the previous example. After the boost in the x direction by $\gamma_1 = 5/4$ (i.e. $\beta_1 = 3/5$) it has coordinates

$$(ct', x', y') = (0,0,0), (0,0,1), (\frac{-3}{4}, \frac{5}{4}, 0), (\frac{-3}{4}, \frac{5}{4}, 1).$$

Now make a boost with $\gamma_2 = 5/3$ (and $\beta_2 = 4/5$) in the positive y direction. The new coordinates are respectively

$$(ct'', x'', y'') = (0,0,0), (-\frac{4}{3}, 0, \frac{5}{3}), (-\frac{5}{4}, \frac{5}{4}, 1), (-\frac{31}{12}, \frac{5}{4}, \frac{8}{3}).$$

These coordinates can not be obtained from those of the original square by a single Lorentz boost in any direction. This may be seen by considering fig. 8.4. If the resultant transformation was a single boost then, its inverse would return the parallelopiped in fig. 8.4 to the original square. However, it is clear that any boost which returns the top right corner to its original position, must displace the top left corner away from the y-axis, in the negative x direction. In fact, if we apply a boost using $\boldsymbol{\beta} = (-4/5, -9/25, 0)$, which corresponds to $\gamma = 25/12$, then all four corners are returned to $t = 0$ but are rotated about the origin (from their positions in the original square), via a pure rotation in the xy plane, by the angle whose cosine is 35/37 and whose sine is 12/37. Thus the coordinates of the corners of the square are (by use of Eq. 8.21)

$$(x, y) = (0,0), (\frac{-12}{37}, \frac{35}{37}), (\frac{35}{37}, \frac{12}{37}), (\frac{23}{37}, \frac{47}{37}).$$

This example illustrates the general result that two successive boosts in different directions are equivalent to a pure rotation followed by a single boost. The rotation involved is called a *Wigner rotation.*

8.4 Poincaré symmetry

The combination of rotations and Lorentz boosts is known as Lorentz symmetry. When combined with translations in space and time we have Poincaré symmetry. The translations in space and time are sometimes called inhomogeneous Lorentz transformations. Poincaré symmetry is thus a basic property of (local) inertial frames and lies at the heart of the Invariance Postulate on which physics is based.

If we are dealing with a restricted inertial frame, then symmetry under boost transformations only applies in one or two dimensions and we thus have restricted Poincaré symmetry (in one or two dimensions). Note that in one space dimension, rotations can not exist.

The Poincaré transformations are all what is called *linear*. This is somewhat of a technical term but relates to the fact that straight lines in one coordinate system map to straight lines in the transformed system. That this is true for Lorentz boosts can be traced to the homogeneity of space and time and the fact that any two inertial frames move at constant velocity with respect to each other (corollary 1.1).

Review

What is the difference between a passive and an active transformation?

Write down the Lorentz transformations for boosts along one of the space axes.

Write down the inverse boost transformations from the new frame to the old.

Sketch a spacetime diagram for frame S showing the path of a light ray and the coordinate axes for another frame S' moving in the positive x direction. Indicate the slope of the S' axes.

Repeat the previous question but this time draw a spacetime diagram for frame S' showing the coordinate axes of S.

The coordinates in a frame S are denoted by (ct, x, y, z) and those in a frame S' moving with relative velocity v by (ct', x', y', z'). In a plot of ct versus any spatial co-ordinate x, y or z a point with spatial coordinates $(x', y', z') = (0, 0, 0)$ in S' will describe which of the following?

- a hyperbola
- the point $(x, y, z) = (0, 0, 0)$

- a parabola
- a straight line

What is rapidity?

What is a Wigner rotation?

What is Poincaré symmetry?

Questions

1. Do the Lorentz transformations apply to an observer travelling faster than the speed of light?

2. The Poincaré transformations apply to inertial frames. Is it feasible to consider transformations to frames which are not inertial? How would you establish a coordinate system in such a frame?

Problems

1. Two twins pilot identical rocket ships containing an identical amount of fuel. They start out at rest in Earth's frame but located 2 ly apart on a straight line leading away from Earth. The two ships start out at the same time in Earth's frame, undergoing identical accelerations away from Earth until their fuel runs out. Since they undergo identical accelerations, they both age identical amounts. However, when the fuel of both has run out, they find that they are not the same age. Explain with the aid of a spacetime diagram and the Lorentz transformations.

2. Show that if large distances are involved then the low velocity limit of the Lorentz time transformation is

$$t' = t - \beta\frac{x}{c}. \tag{8.29}$$

   Consider two clocks separated by 4.007×10^7 m in a frame in which they are synchronized. In another frame moving with speed 465 m/s with respect to the first, show that the clock times differ by 207.3 ns. (The distance here corresponds to the circumference of the Earth and the relative velocity of the frames corresponds to the uniform rotational speed of a clock on the Earth's surface as seen by an observer at rest with respect to Earth's center. This time difference is at the heart of the Sagnac effect in rotating reference frames.)

3. It is found experimentally that when light is reflected from a mirror at rest then the angles subtended by the incident and reflected beams with respect to a perpendicular to the mirror are equal. In the simple arrangement shown in fig. 8.5, where the light is emitted and detected the same distance h from the mirror, this corresponds to the light arriving at the detector having been reflected from a point midway between source and detector.

   (a) Consider the travel time of the reflected light from source to detector. Show that the observed path is the one for which the total travel time is a minimum. This result is known as Fermat's Principle.

   (b) Make an arbitrary Lorentz boost to another inertial frame. Show that Fermat's Principle still holds.

4. In 1609, after studying the astronomical data of Tycho Brahe, Johannes Kepler announced the first two of his three laws of planetary motion. He asserted that

   - the planets move about the Sun in elliptical orbits with the Sun at one focus of the ellipse;
   - during equal time intervals, the line joining the Sun and a planet sweeps out equal areas.

   Consider the second of these assertions in the special case of a circular orbit, of radius R. Let the orbit be in the xy plane with the Sun at the origin. Deduce that in equal time intervals, a planet must move equal distances around this circular orbit.

   Consider the four equally spaced points

   $$(x, y) = (\frac{R}{\sqrt{2}}, \frac{R}{\sqrt{2}}),\ (-\frac{R}{\sqrt{2}}, \frac{R}{\sqrt{2}}),\ (-\frac{R}{\sqrt{2}}, -\frac{R}{\sqrt{2}}),\ (\frac{R}{\sqrt{2}}, -\frac{R}{\sqrt{2}})$$

   i.e. points at $45°$, $135°$, $225°$ and $315°$. Prove that in a frame moving along the x axis with speed v, the four areas corresponding to the time intervals between these points remain equal but that the time intervals themselves do not.

Is Kepler's "second law" valid in this new frame?

5. Consider the twin paradox. If, in the frame of the stay-at-home twin, the traveller travels a distance D, at β times the speed of light, the total time measured by the stay-at-home twin for the journey is clearly $ct = 2D/\beta$. Show that in the frame of the outward bound traveller, the outward journey begins and ends at $x' = 0$ and takes time

$$ct' = \frac{D}{\gamma\beta}.$$

Next, show that in the frame of the homeward bound traveller, the journey home begins and ends at $x'' = \gamma 2D$ while the time at which he begins returning is

$$ct'' = \gamma\left(\frac{D}{\beta} + \beta D\right)$$

and he gets home at

$$ct'' = \frac{\gamma 2D}{\beta}.$$

Hence confirm that the return journey takes the same time for the returning traveller as the outward traveller measured for the outward journey. However, show also that in the frame of the returning traveller the outward journey took a time

$$ct'' = \frac{D}{\gamma\beta} + 2\gamma\beta D,$$

i.e. $2\gamma\beta D$ more than that measured by the outward bound traveller. Thus show that the total time elapsed for the stay-at-home twin, during the journey, as measured by the returning traveller, is

$$c\Delta t'' = \gamma\frac{2D}{\beta}.$$

Infer, via an inverse transformation [be careful here!], that the total time elapsed on a clock left at home, is $2D/\beta$; exactly as measured by the stay-at-home twin. (The origin for all frames is to be taken at the departure point.)

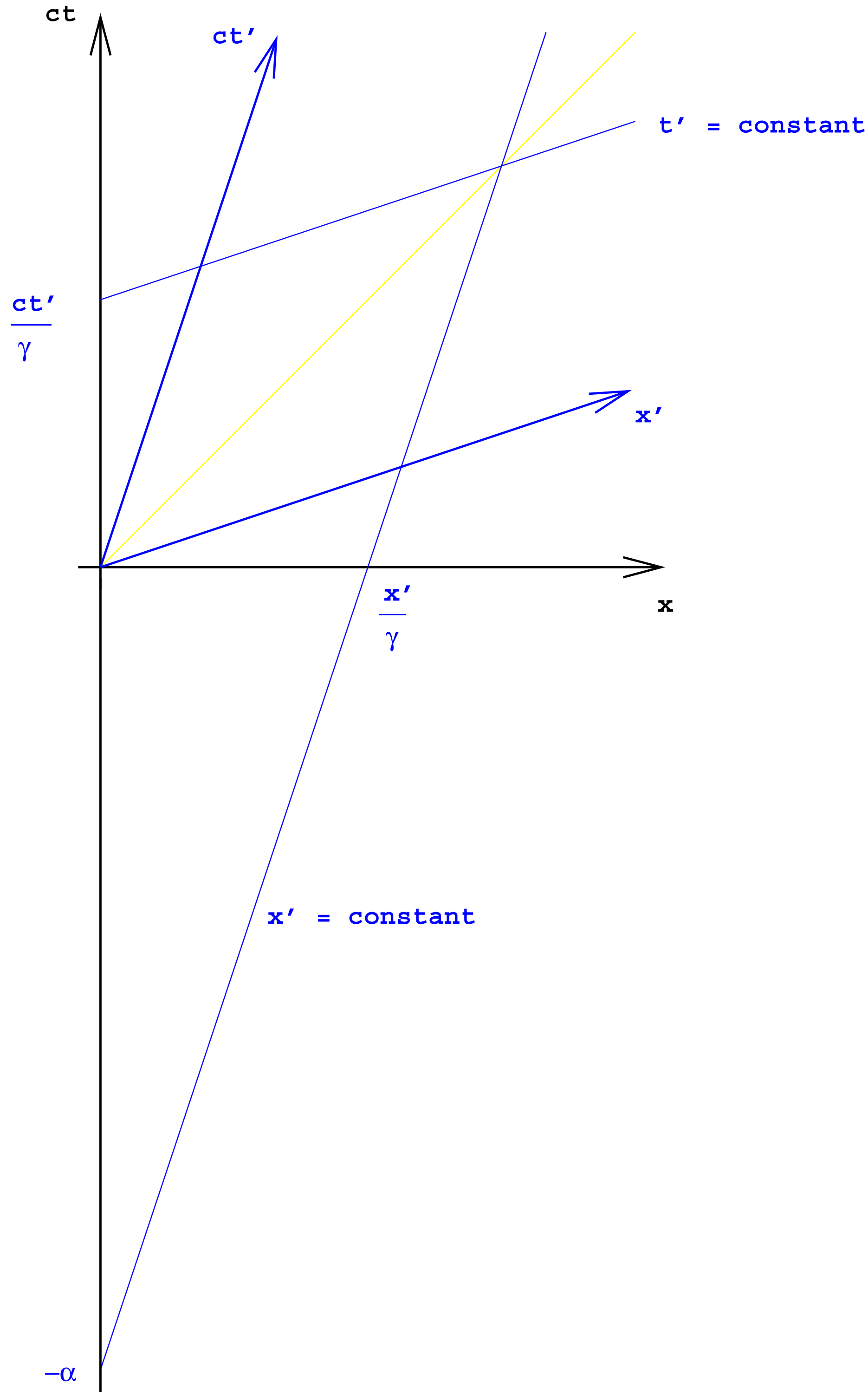

Figure 8.2: The coordinate system of a measuring rod as it appears in a frame in which the rod is moving with speed v.

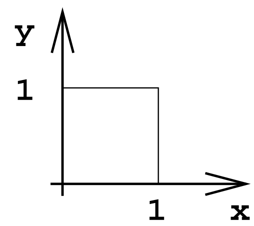

Figure 8.3: A square in the xy plane.

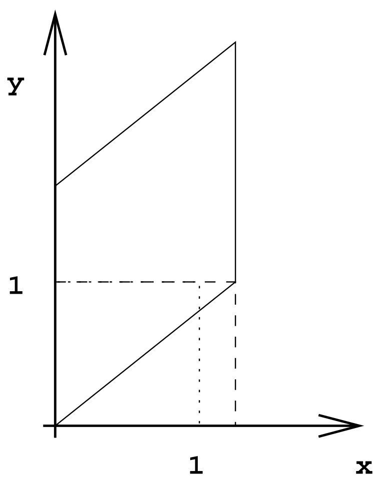

Figure 8.4: A square (dotted lines) is first boosted by $\beta_1 = 3/5$ in the x direction, leading to the dashed rectangle, and then is boosted again by $\beta_2 = 4/5$ in the y direction, leading to the parallelopiped with solid lines.

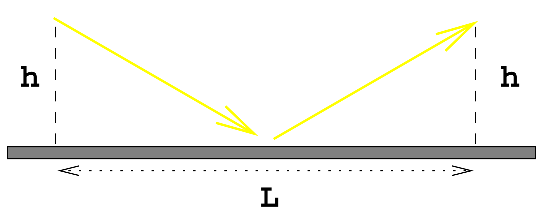

Figure 8.5: Light reflected from a mirror.

Chapter 9

The Cosmic Speed Limit

9.1 Causality

By *causality* we mean the physical principle that effects follow from their causes (and those causes must therefore precede their effects). Thus, for example, if a boy swings a bat, hitting a ball, which then flies through the air and breaks a window, we say that the breaking of the window was caused by the hitting of the ball. If a firing squad shoots a condemned prisoner, that prisoner dies as a consequence of the firing squad pulling the triggers on their rifles, thus causing bullets to be projected towards the prisoner and hitting him. If a glass is dropped on the floor and breaks, we legitimately say that the breaking was a consequential effect of the glass being dropped. If a male and a female have sex, then the birth of a baby nine months later is caused by that act. Even primitive peoples with limited knowledge of biology and the intricacies of conception understand that having sex is what makes women pregnant and the birth of a child results from that.

Thus cause and effect are events with an identifiable relationship between them such that it is indisputable that the effect (event 2) is a consequence of the cause (event 1). It follows that all inertial observers must agree on causal relationships, and consequently they must also agree on the order in which cause and effect occur. Otherwise, there would be a distinction between inertial frames in which boys hitting balls caused broken windows and frames in which broken windows caused boys to hit balls. This would violate the Invariance Postulate and is therefore not permitted in our physics.

9.1.1 Time travel

It is to be noted that causality places restrictions on time travel to the past. We saw in our discussion of time dilation that travel to the future was possible, in principle, by travelling at very high speeds. However, travel to the past presents problems. The classic example is of a person travelling to a time in their past and killing a parent, grandparent, or other ancestor before they were conceived. But if their ancestor dies before the person is conceived then that person will never be born and therefore could not have travelled to the past to do the killing.

The possibility of causal loops with contradictions built in leads most scientists to reject the possibility of time travel to the past. Though some ingenious schemes have been devised for removing such inconsistencies they involve modifications of the way the universe and its inhabitants are believed to operate and are not generally accepted.

9.1.2 The increasing entropy of the universe

The study of macroscopic collections of molecules, such as occur in gases, involves an enormous number of particles, of the order of Avogadro's number, i.e. about 10^{26}. It is quite impossible to follow the motion of every molecule individually and so one instead deals with the distribution of molecule positions and velocities, etc., and their averages. If f_ν is the probability of finding molecules in state ν then by definition of probability,

$$0 \leq f_\nu \leq 1. \tag{9.1}$$

(It is essentially just the fraction of molecules in the state ν, which for a homogeneous system could be specified by the position and velocity of molecules.) For a system in equilibrium, the distribution f_ν will remain constant (for each ν) but in general it will change due to the collisions of molecules with each other and the walls of their container. We write this change as

$$\frac{df_\nu}{dt} = Q_\nu, \tag{9.2}$$

where Q_ν is known as the collision term and represents the net gain of molecules in state ν as a result of molecular interactions. The equation is known as Boltzmann's equation. The form of Q_ν is of some interest. It is basically the difference between the likelihood that molecules in other states are scattered to state ν (thus increasing the number in state ν) minus the likelihood that molecules in state ν are scattered to other states. If $T_{\nu\nu'}$ is the probability that a *given* molecule in state ν, in the gas, makes a transition to state ν' then we may write

$$Q_\nu = \sum_{\nu'} \left[f_{\nu'} T_{\nu'\nu} - f_\nu T_{\nu\nu'} \right]. \tag{9.3}$$

It is useful to introduce the *entropy*, S, of the system by defining

$$S = -k \sum_\nu f_\nu \log f_\nu \tag{9.4}$$

where k is a constant of proportionality known as Boltzmann's constant. If all molecules belong to a single state then $f_\nu = 1$ for that state and 0 for all other states. Hence $S = 0$. However, if the molecules are distributed one per state then $f_\nu = 1/N$ for all occupied states, where N is the total number of molecules. In this case $S = k \log N$.

Exercise 9.1 *Confirm the stated values of S for these cases and show that S is always greater than or equal to zero.*

If a set of N coins is lying on a table after having been shaken, what is the entropy if they are all lying with heads up? What is the entropy if the number of heads is equal to the number of tails (which is the most likely result if they fall head up or tail up with equal probability after having been shaken)?

Thus entropy may be regarded as a measure of the degree of disorder in the system — highest order corresponding to molecules being in the fewest possible number of states. Since information is associated with patterns, rather than random noise, entropy is also associated with the information content of a system and this in turn leads to applications in the field of data compression.

Now, if one considers the time rate of change of S one finds that

$$\frac{dS}{dt} = -k \sum_\nu Q_\nu \log f_\nu. \tag{9.5}$$

In general, the form of this equation places little constraint on the derivative of S. However, if we insert the specific form of Q_ν presented in Eq. 9.3 we obtain, after a change of dummy summation indices:

$$\begin{aligned}
\frac{dS}{dt} &= -k\sum_{\nu}\sum_{\nu'}\left[f_{\nu'}T_{\nu'\nu} - f_{\nu}T_{\nu\nu'}\right]\log f_{\nu} \\
&= +k\sum_{\nu}\sum_{\nu'}\left[f_{\nu'}T_{\nu'\nu} - f_{\nu}T_{\nu\nu'}\right]\log f_{\nu'} \\
&= +k\frac{1}{2}\sum_{\nu}\sum_{\nu'}\left[f_{\nu}T_{\nu\nu'} - f_{\nu'}T_{\nu'\nu}\right]\log\frac{f_{\nu}}{f_{\nu'}}. \qquad (9.6)
\end{aligned}$$

Under some circumstances

$$T_{\nu\nu'} = T_{\nu'\nu}. \qquad (9.7)$$

This can be thought of as a time reversal symmetry for one molecule interacting with the rest of the gas. Time reversal symmetry is not one of our fundamental symmetry postulates and is not true in general even on a microscopic scale. However, it is valid when a pair of classical molecules interact and from this can be shown to be valid for a gas comprising a very large number of molecules under the further assumption that the molecular motions are not correlated and interactions only occur between pairs of molecules, and provided also that certain boundary conditions are met at the walls. (A transfer of heat from the molecules to the walls, such as occurs inside a refrigerator, would be one circumstance when these boundary conditions were not satisfied.) When this symmetry applies we can write

$$\frac{dS}{dt} = +k\frac{1}{2}\sum_{\nu}\sum_{\nu'}\left(f_{\nu} - f_{\nu'}\right)T_{\nu\nu'}\log\frac{f_{\nu}}{f_{\nu'}}. \qquad (9.8)$$

This is always either positive or zero (because the log term is negative when $f_{\nu} < f_{\nu'}$ and positive otherwise so that every term in the summation is either positive or zero). The zero case occurs for a system in equilibrium (since each Q_{ν} in Eq. 9.5 is then zero). Thus the entropy always increases with time until it reaches a maximum. This result is often referred to as the second law of thermodynamics.

How can this be if there is nothing special about the direction of time? The answer is rooted in causality. The collision term specifies how the distribution of molecules changes as a *result* of collisions. The overall sign in Eq. 9.3 is dictated by this; it arises because Q_{ν} depends on the initial configuration of states and the probabilities for *changes* in individual states. Even if individual collisions exhibit time reversal symmetry, the overall system need not. Thus one cannot simply run time backwards (i.e. reverse all of the motions) and reverse the direction of entropy change. The causal nature of the collision term prohibits this and dictates that the forward direction of time is special.

9.2 The ultimate speed

We have earlier seen that events that are simultaneous in one inertial frame will not be in another. But can the time ordering of events be changed by observing from a different frame?

Consider a pair of events, 1 and 2, in any given inertial frame. Let them have coordinates (ct_1, x_1, y_1, z_1) and (ct_2, x_2, y_2, z_2) respectively. In a frame moving along the x axis the new time coordinates are:

$$\begin{aligned} ct'_1 &= \gamma(ct_1 - \beta x_1) \\ ct'_2 &= \gamma(ct_2 - \beta x_2). \end{aligned}$$

Thus, if the events are separated by $\Delta t = t_2 - t_1 > 0$ and $\Delta x = x_2 - x_1$ in the first frame, then, in the new frame we have

$$c\Delta t' = \gamma(c\Delta t - \beta\Delta x). \tag{9.9}$$

Strictly, the Lorentz transformations have only been derived for $\beta \leq 1$. Even so, it is clear that by making Δx large enough, $c\Delta t$ can still be made negative (even for quite small values of β).

What does this mean? Suppose that events 1 and 2 were causally related. Then some information or signal must have propagated between them. (In the above example of the boy hitting the ball and breaking the window, the signal is carried by the ball.) This signal must therefore propagate with speed

$$u = \Delta x/\Delta t. \tag{9.10}$$

However, if the time ordering is altered by the Lorentz transformation (with $\beta \leq 1$) then

$$\beta\Delta x > c\Delta t \tag{9.11}$$

which implies that

$$u > c/\beta \geq c. \tag{9.12}$$

Thus the signal would have to have travelled faster than the speed of light (in vacuum) in the initial observer's frame in order for these events to be causally related.

We conclude therefore that the time ordering of causally related events is not altered by the Lorentz transformations, *provided* that all signals in the universe are restricted to speeds less than c. If this condition is not satisfied then causality can be violated and physics, as we know it, breaks down. The assumed validity of the Invariance Postulate, and in particular the Principle of Relativity, has thus led us to the following result.

Theorem 1 (Ultimate Speed) *No signal may be transmitted at a speed greater than c.*

Since any object can carry a signal, this means that nothing can travel faster than light!

9.3 Spacetime regions

Consider the path of a light ray on a ct versus x spacetime diagram. A ray emitted by an observer at the origin travels either in the positive or negative x direction and the trajectories are symmetrical lines at 45° to the x axis. Similarly, with light arriving at the origin from times in the past. If we now add a y dimension then the possible trajectories of the light rays form a pair of cones, known as the future light-cone and the past light-cone. See fig. 9.1.

Points lying inside the future light-cone have $(ct)^2 - \mathbf{x}^2 > 0$ and thus can be communicated with by the origin. Therefore events at those points can be influenced by an event at the origin. Those points can also, in principle, be reached by an observer starting at the origin and so are said to lie in his future. Similarly, points lying inside the past light-cone have a positive spacetime interval and can communicate with and influence an observer at the origin. They are said to lie in that observer's past.

In contrast, events lying outside of either light-cone can not communicate with the origin. They are completely unreachable, though events in those regions may still influence the future of an observer at the origin.

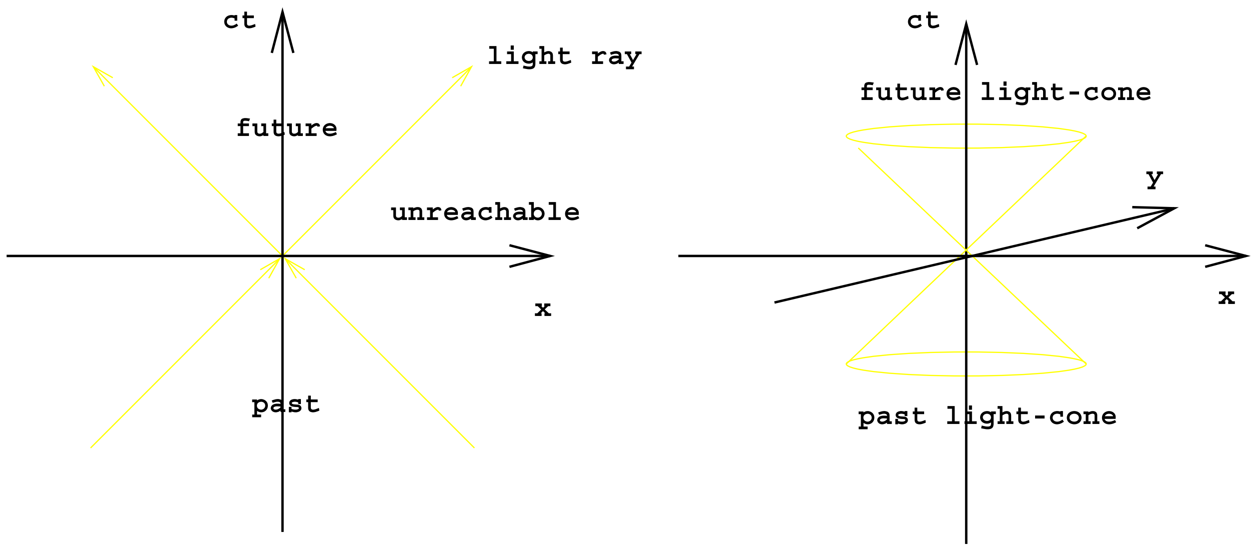

Figure 9.1: The light-cone in one and two space dimensions.

When we move to three space dimensions, the light-cone becomes an expanding sphere. It is still common though to refer to it as the light-cone and this usage also applies in just one space dimension.

It should of course be clear that each point in spacetime has a light-cone associated with it. Light-cones are not limited to emanating from the origin. This leads us to classify general spacetime regions according as to whether or not communication is possible.

Let $c\Delta t = c(t_2 - t_1)$ and $\Delta\mathbf{x} = (x_2 - x_1, y_2 - y_1, z_2 - z_1)$ be the time and space separations of two events. We say that the spacetime interval $(c\Delta t)^2 - (\Delta\mathbf{x})^2$ is

$$\begin{array}{rll} \textit{timelike} & \text{if} & (c\Delta t)^2 \text{ dominates} \\ \textit{spacelike} & \text{if} & \Delta\mathbf{x} \text{ dominates} \\ \textit{lightlike} & \text{if} & (c\Delta t)^2 = (\Delta\mathbf{x})^2. \end{array}$$

Because some authors define the spacetime interval to be $(\Delta\mathbf{x})^2 - (c\Delta t)^2$, one should understand the definitions of these terms as given above and not try and learn them in terms of whether the spacetime interval is positive or negative.

With this terminology, we can state the results of the last section as

- Events with a spacelike separation can not communicate with each other and their time ordering depends on the inertial frame of the observer.

- Events with a timelike separation can be connected by a signal travelling at a speed less than that of light and their time ordering is the same for all observers.

- Events with a lightlike separation can only be connected by light propagating in vacuum and their time ordering is the same for all inertial observers, though the time separation will approach zero for observers in the frame of the light signal connecting the events.

9.4 Rigidity

The inability of objects with spacelike separations to communicate has important implications for our concepts of rigidity. Consider, for example, what happens when one pushes a very long iron rod. One's perception may be that the far end instantly starts to move. However, this cannot be true because the end that is pushed and the far end have a spacelike separation. There is no way that the far end can know to move until a signal has reached it. This signal is in fact a shock wave that propagates down the rod at the speed of sound in the rod. That speed increases as the rigidity of the rod increases but can never exceed c. The situation is just like pushing a gelatine pudding, or a long spring. It takes time for the impulse to travel through the pudding, or spring. An iron rod is no different. The speed of the impulse may be much greater but it is still finite and less than c.

A similar situation occurs if we swing an iron rod. It may seem that if the rod were long enough and if we were strong enough to swing it then the far end would rotate so that its speed was as high as we wanted. Certainly this would be true if the rod were perfectly rigid because the velocity of the far end is given by $v = r\omega$ where r is the rod length and ω the angular velocity of rotation. Even for ω small, if r is large enough then the far end could move at any speed desired, including faster than light. The catch here is that the rod is not rigid. If one starts to swing it in an arc, the far end does not start moving instantly. The separation of the far end and the end held in one's hand is spacelike and the far end does not know to start moving until a signal reaches it. It is like swinging a flexible plastic rod. The rod bends as it is swung, with the far end lagging behind the rest. A signal must be propagated down the rod, at the speed of a transverse wave in the rod, and that speed is finite and less than c. It is easy to see that no matter how long the rod, the far end can not be forced to move faster than c by swinging the rod from the near end. The only difference between the iron rod and plastic rod is the degree of bending; it must be present in both.

The conclusion to be reached from these examples is that there is no such thing in relativity theory as a perfectly rigid object. The following example underscores this very well.

Example Ultra Undertakers deliver a coffin to the graveyard by sliding it along the ground at such a high velocity that the Lorentz contraction factor is $\gamma = 5$. The gravediggers have dug a hole for the coffin with the same proper length as the proper length of the coffin itself. Thus a snug fit is possible when the coffin is at rest. In the frame of the gravediggers, the coffin is Lorentz contracted by $1/5$ and so the coffin readily falls in. However, in the frame of the undertakers, it is the grave which is length contracted and so the coffin will surely not fall in. Both views can not be correct though; the coffin will either get buried or it will not.

This problem can be correctly analysed by ignoring any tipping of the coffin as its front edge passes over the hole. If we wish, this can be arranged via a trap door setup. Also, the constant acceleration of gravity as the coffin falls is inessential — a similar situation could be envisaged with objects being accelerated in a downward blast of air, or an electric field.

As so often happens, the resolution of this paradox rests with simultaneity. If we say that the coffin begins to fall in the gravedigger's frame at $t = 0$ (i.e. the instant the rear of the coffin passes over the edge of the grave) then in the undertaker's frame, the rear will also start to fall at $t' = 0$ but the front will start to fall at an earlier time. In fact the coffin must *flow* into the hole!

Review

What is causality?

Is the special nature of the forward direction of time responsible for causality or is it the result of causality? What physical principle is associated with this special nature of the forward direction of time?

What is meant by a timelike, lightlike and spacelike spacetime interval?

What is the fastest speed an object can travel at and why?

Can an object be perfectly rigid? Why?

Questions

1. In the example of the sex act being causally related to childbirth, what carries the signal?

2. Is there a causal connection between an alignment of the planets in the sky and the fate of governments?

3. Does the outcome of a sports game depend on your presence as a spectator in the viewing stands?

4. Could the contradictory causal loops associated with time travel be eliminated by denying people a free will? Is this reasonable? Would multiple universes, differing only in their histories, be an acceptable way out?

5. The state of maximum entropy can be shown to be the most likely configuration. It follows therefore that if a system is not in its most likely configuration, then after a change has taken place it is more likely that entropy has increased rather than decreased. Is this purely probabilistic state of affairs equivalent to the law of increasing entropy that constitutes the second law of thermodynamics?

   Consider the following example. A sizeable number of coins are lying on a table, most of them head up. The table is shaken hard so that the coins are tossed about. When they come to rest again it is most likely that the number of heads and tails facing up is morely evenly distributed (i.e. the coins are in a state of increased entropy) but there is a non-zero probability that the number of heads facing up will increase (i.e. a decrease in entropy). Does this mean that the law of increasing entropy is only approximately true or are there reasons why the law is not applicable to this particular situation?

6. The derivation of the theorem on the ultimate speed hinges on the Lorentz boost transformation for time. Consider this equation in the non-relativistic limit. Can the time ordering of events depend on the reference frame in that limit? Does the low-velocity limit imply an ultimate speed?

7. Could the spot of light on an oscilloscope screen, sweep across the screen at a speed greater than c?

Problems

1. A power surge at mission control in Houston causes a malfunction of vital equipment. Exactly one second later, a piece of equipment in a landing craft on the Moon, 3.8×10^8 m away, explodes. Could the two incidents be related? Explain.

2. (a) A ruler is inclined at a small angle θ to a straight edge. The ruler is moving with speed v in a direction perpendicular to the edge. Consider the point of intersection of ruler and edge. Show that this point moves along the straight edge at a speed

$$v_p = \frac{v}{\tan\theta}.$$

   Can this be faster than c? Whatever your answer, reconcile it with the theorem on the ultimate speed.

   (b) A pair of scissors is closed. Could the point at which the blades meet move along the blades at a speed faster than c? Explain fully, using both physical principles and either analogy or contrast with the situation in the first part of this problem.

3. Consider a situation in which a person goes back in time and kills a grandparent.

   (a) Sketch a spacetime diagram showing the events:

      - A grandparent begets parent,

- B parent begets child,
- C child kills grandparent,

and sketch worldlines of parent, child and grandparent connecting these events.

(b) Show that a violation of causality exists if and only if the grandparent travels from A to C at an average speed faster than that of light (and note that this is independent of the worldlines of parent and child except insofar as their worldlines combine to also join events A and C).

(c) Write a few sentences discussing whether or not a violation of causality requires a round trip signal, or a closed loop in a spacetime diagram. Does this have any bearing on the derivation of the theorem on the ultimate speed (which only uses a one-way signal trip)?

4. Speedy Gonzalez, the super fast mouse, is zipping along at close to light speed. Bad Johnny Brat, a very naughty boy, is determined to give Speedy a fright. He waits, with two meat cleavers held aloft and as Speedy passes brings them down simultaneously, 4 inches apart, so that the front one falls just before Speedy's nose (his foremost part) arrives. In his rest frame, Speedy is 5 inches long, but Bad Johnny Brat sees Speedy Lorentz contracted to 3 inches and so expects his second cleaver to miss Speedy's tail and thus he merely intends to frighten Speedy. On the other hand Speedy sees the separation between the meat cleavers as Lorentz contracted to less than 4 inches and so is most alarmed. Fearing that he is about to meet his end he sends a radio message to the Society for the Prevention of Cruelty to Animals, alerting them as to the identity of his murderer.

   (a) How fast is Speedy travelling?

   (b) In Speedy's frame, how far apart are the meat cleavers?

   (c) Sketch a spacetime diagram in Speedy's frame showing clearly the world lines of the two meat cleavers and those of Speedy's nose and tail as well as the light cone. Mark the events: fall of cleaver in front of nose as event 1 and the fall of the second cleaver as event 2.

(d) Does Speedy get sliced or not? Explain clearly. If you think he does not, reconcile the 4 inch separation of the cleavers with the 5 inch rest frame length of Speedy.

5. A spider has discovered a home in a small crevasse at the bottom of a bolt hole in some very strong steel. Garfield, Jim Davis' spider-hating cat, sees the spider and fires a bolt at relativistic speed into the hole. The proper length of the bolt is just short enough to allow the spider room at the bottom of the bolt hole in the hole's rest frame. The bolt also has a very strong head which prevents it penetrating the hole any further. However, at the speed that Garfield fires the bolt the hole appears to be considerably contracted in the bolt's frame. The spider should therefore get squashed if he does not retreat into his crevasse. But in the spider's frame, it is the bolt that is contracted and there would appear to be plenty of room for the spider. Garfield waits with anticipation. Will the spider get squashed or not?

6. A conducting slider closes the circuit in fig. 9.2, and the light bulb shines. The slider is of proper length L_0 and passes the bent section of wire whose dimensions are such that in the frame of the wire the Lorentz contracted slider leaves a break in the circuit. Thus the bulb should go off as the slider passes. However, in the frame of the slider

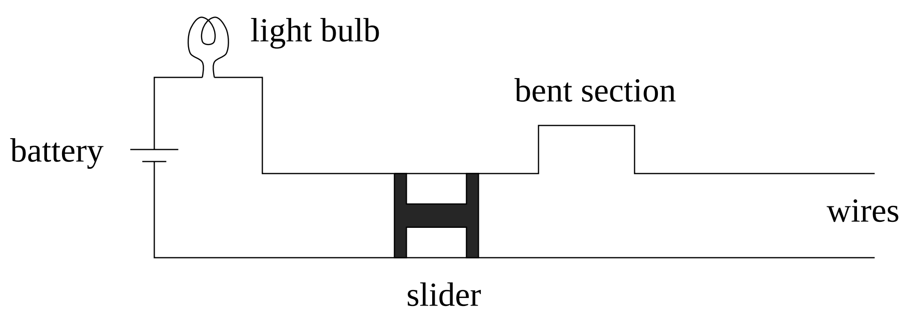

Figure 9.2: Flickering light bulb paradox

it is the length of the bent section that is Lorentz contracted and thus a closed circuit exists at all times. Therefore, the bulb will not go off. Both descriptions can not be correct though; the bulb will either go off briefly (i.e. flicker) or it will not. Resolve this paradox.

7. A train of proper length L_0, travelling at very high speed, enters a tunnel of just slightly greater proper length. The far end of the tunnel happens to be blocked by a huge landslide. A railway worker sees the train Lorentz contracted so it easily fits inside the tunnel. Spiteful from having been sacked, the railway worker triggers an explosion at the open end at an instant when the train is inside the tunnel. The explosion causes another huge landslide which traps the train inside. However, in the train frame it is the tunnel which is Lorentz contracted and so the train does not fit inside the tunnel and cannot be trapped inside by the railway worker. But the train is either trapped or it is not. Is the railway worker able to carry out his evil plan?

8. Quasars are exceptionally bright objects located at the farthest reaches of the universe. They are sometimes observed to eject luminous matter. In quite a few cases, the rate of increase of the angular separation of ejected matter and quasar, when combined with the assumed large distance to the quasar, makes it seem as though the matter is ejected at a speed faster than c. Let the matter be ejected with speed v at an angle θ to the line of sight. Recalling the derivation of the transverse doppler effect (in chapter 7), show that the apparent (i.e projected) rate of separation across the field of view is

$$v_{\text{app}} = \frac{v \sin\theta}{1 - \beta\cos\theta}.$$

Show that the maximum occurs for $\cos\theta = \beta$ and that v_{app} can exceed c. How large must v be for this to happen? Is there any contradiction with the theorem on the ultimate speed?

9. A Zen master is accosted by bandits on a mountain road. One of the bandits fires an arrow of proper length 80 cm at the Zen master's head. The Zen master holds his arm out and miraculously catches the arrow, which is flying at $0.6c$, by the feathered tail end as it passes, stopping it in just 1 cm! In the frame of the Zen master (and the bandit) the arrow is contracted and is shorter than the Zen master's outstretched arm, which is of proper length 85 cm from shoulder to wrist. The Zen master should therefore be unharmed. However, in the frame of the arrow it is the Zen master's arm which is contracted and he should be killed. Which version is correct?

Chapter 10

Velocity Transformations

The velocity of an object is a vector quantity with components given by $\mathbf{u} = (u_x, u_y, u_z)$. By definition, the velocity is

$$\mathbf{u} = \frac{d\mathbf{x}}{dt} = \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right). \tag{10.1}$$

Thus the components are

$$u_x = \frac{dx}{dt} \equiv \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} \tag{10.2}$$

$$u_y = \frac{dx}{dt} \equiv \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t} \tag{10.3}$$

$$u_z = \frac{dx}{dt} \equiv \lim_{\Delta t \to 0} \frac{\Delta z}{\Delta t}. \tag{10.4}$$

10.1 Velocities in different frames

In any other frame, the components of velocity must have the same form. For example,

$$u'_x = \frac{dx'}{dt'}. \tag{10.5}$$

We seek to relate the velocity in the new frame to that in the old. This is easily accomplished via the Lorentz transformations.

Let the new frame be moving in the positive x direction with velocity v, then Δx, Δt etc. transform according to Eqs. 8.12-8.15, since they are just differences of two variables which transform linearly, i.e.

$$\begin{aligned} c\Delta t' &= \gamma(c\Delta t - \beta\Delta x) \\ \Delta x' &= \gamma(\Delta x - \beta c\Delta t) \\ \Delta y' &= \Delta y \\ \Delta z' &= \Delta z. \end{aligned}$$

Since the differentials, dx, dt, etc., are just the infinitesimal limits of Δx, Δt etc., they have the same transformations:

$$\begin{aligned} cdt' &= \gamma(cdt - \beta dx) && (10.6) \\ dx' &= \gamma(dx - \beta cdt) && (10.7) \\ dy' &= dy && (10.8) \\ dz' &= dz. && (10.9) \end{aligned}$$

Thus

$$\begin{aligned} u_x' = \frac{dx'}{dt'} &= \frac{\gamma(dx - \beta cdt)}{\gamma(dt - \frac{\beta}{c}dx)} \\ &= \frac{\frac{dx}{dt} - v}{1 - \frac{v}{c^2}\frac{dx}{dt}} \\ &= \frac{u_x - v}{1 - \frac{vu_x}{c^2}}. && (10.10) \end{aligned}$$

Also, we have

$$\begin{aligned} u_y' = \frac{dy'}{dt'} &= \frac{dy}{\gamma(dt - \frac{\beta}{c}dx)} \\ &= \frac{\frac{dy}{dt}}{\gamma(1 - \frac{v}{c^2}\frac{dx}{dt})} \\ &= \frac{u_y}{\gamma(1 - \frac{vu_x}{c^2})}. && (10.11) \end{aligned}$$

Similarly,

$$u_z' = \frac{u_z}{\gamma(1 - \frac{vu_x}{c^2})}. \qquad (10.12)$$

As with the Lorentz transformations of the coordinates, the inverse transformations are given by changing the sign of v.

Notice that the change in the transverse components of velocity stems from the transformation in the time variable. This transformation of the time in the denominators is responsible for the complicated looking nature of the velocity transformations.

It is useful to consider some special cases. Firstly, consider the low velocity (Galilean) limit when $u_x \ll c$ and $v \ll c$. In that case $\gamma \approx 1$ and $1 - vu_x/c^2 \approx 1$. Thus one obtains the Galilean velocity addition transformations:

$$\begin{aligned} u'_x &= u_x - v \\ u'_y &= u_y \\ u'_z &= u_z. \end{aligned} \tag{10.13}$$

These transformations are commonly used in (pre-relativity) Newtonian physics but are recognized here as only low velocity approximations. Nevertheless, it is satisfying that this low velocity approximation is contained within the more general velocity transformations.

Next we consider the high velocity limit when the object has a high velocity in the direction of motion of the new frame, i.e. we consider the limit $u_x \to c$. We find that

$$u'_x \to \frac{c - v}{1 - \frac{vc}{c^2}} = c \tag{10.14}$$

independently of v. This is consistent therefore with our requirement that the speed of light be the same in all inertial frames. Of course the experimental evidence for the independence of light speed on source velocity also constitutes evidence for the validity of the velocity transformations.

Exercise 10.1 *A few quick manipulations would appear to show that in the high velocity limit $u_x \to c$:*

$$u'_y \to u_y \sqrt{\frac{1+\beta}{1-\beta}}.$$

Derive this for yourself. What happens when $\beta \to 1$? Why can this not be correct? What is u_y in this same limit? What went wrong in the derivation?

10.2 Light in a moving medium

Due to repeated absorptions and re-emissions, light travels in media with a reduced speed

$$c_{\text{medium}} = c/n \tag{10.15}$$

where n is known as the *refractive index* of the medium and is necessarily always greater than 1.

Consider now a medium moving with speed v parallel to the direction of light propagation. In the rest frame of the medium, the light moves with velocity $u'_x = c/n$. Thus in the laboratory frame, in which the medium has velocity v, the speed of light is

$$\begin{aligned} u_{\text{light}} &= \frac{u'_x + v}{1 + \frac{vu_x}{c^2}} \\ &= \frac{\frac{c}{n} + v}{1 + \frac{v}{c^2}\frac{c}{n}} \\ &= \frac{c}{n}(1 + \frac{nv}{c})(1 + \frac{v}{nc})^{-1}. \end{aligned} \tag{10.16}$$

In the low velocity limit $v \ll c$ this reduces to

$$\begin{aligned} u_{\text{light}} &= \frac{c}{n}(1 + \frac{nv}{c})(1 - \frac{v}{nc} + \cdots) \\ &= \frac{c}{n}(1 + \frac{nv}{c} - \frac{v}{nc} + \cdots) \\ &\approx \frac{c}{n} + vf \end{aligned} \tag{10.17}$$

where

$$f = 1 - \frac{1}{n^2} \tag{10.18}$$

is known, for historical reasons, as the Fresnel drag coefficient. Fresnel originally introduced this coefficient in 1818 in order to explain the observation by his fellow countryman Arago that the refraction of starlight by glass appeared to take place as if the glass were not moving (through a supposed ether, which it effectively should have been due to the orbital velocity of Earth in the supposed ether).

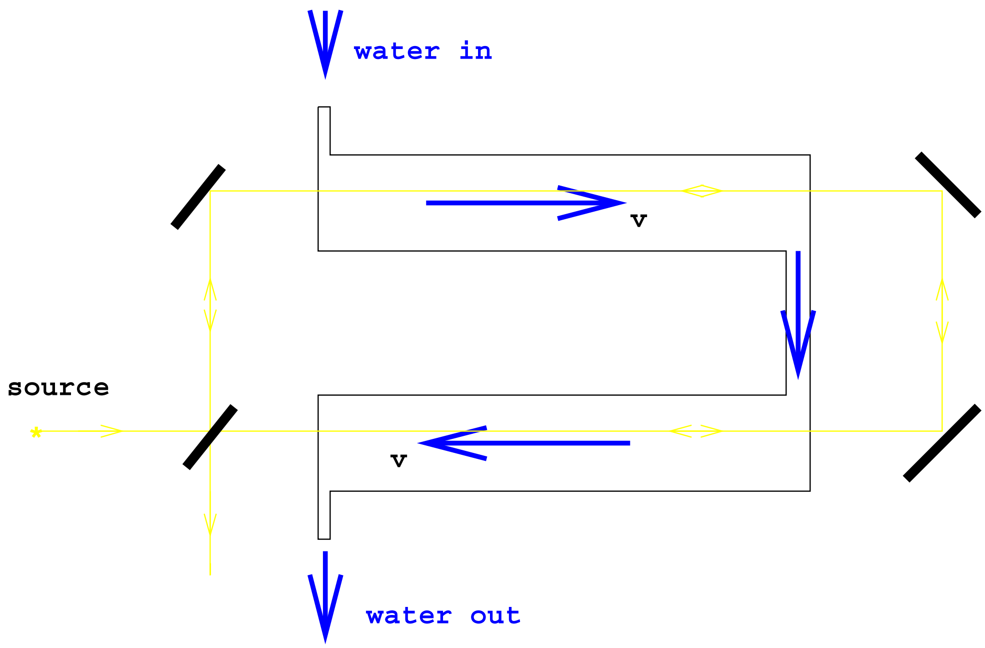

Figure 10.1: Schematic of the Fizeau experiment.

One aspect worth pointing out is that the speed of light in a medium is not the maximum allowed speed in that medium. The maximum speed in a medium is still the velocity of light in vacuum because that maximum comes from causality arguments and the Lorentz transformations rather than being a property of light itself. Indeed high speed cosmic rays quite commonly travel through water at faster than the speed of light in water. Charged particles that travel through a medium at a speed faster than that of light in the medium emit what is known as Cerenkov radiation. This is somewhat like a sonic boom from supersonic aircraft flying through air.

10.2.1 The Fizeau experiment

The Fizeau experiment was of considerable historical importance in the development of relativity theory. In the experiment, light is sent in two directions by a beam splitter. One beam travels upstream through moving water and the other travels downstream through the same water. See fig. 10.1. By a clever arrangement, the two beams are made to recombine. An interference pattern results from the differences in the paths travelled by the two beams and when the water is flowing this fringe pattern shifts from the position it had with the water stationary.

The modern interpretation of this experiment is that light is propagating in a moving medium with speed given by Eq. 10.17 while travelling downstream and with speed given by

$$u_{\text{light}} \approx \frac{c}{n} - vf \tag{10.19}$$

while travelling upstream. Thus the time difference for light travelling the two paths is

$$\begin{aligned}\Delta t &= \frac{2l}{(\frac{c}{n} - fv)} - \frac{2l}{(\frac{c}{n} + fv)} &(10.20)\\ &\approx \frac{4n^2 fvl}{c^2}. &(10.21)\end{aligned}$$

Thus the fringe shift is given by

$$\delta = \frac{c\Delta t}{\lambda} = \frac{4n^2 fvl}{\lambda c}. \tag{10.22}$$

Fizeau's original experiment, showing the fringe shift, was repeated with greater precision by Michelson and Morley in 1886 and later by P. Zeeman and associates in 1914-1922. The influence of the motion of the medium on the propagation of light has thus been verified.

10.3 The headlight effect

Consider a beam of light (say from a flashlight) in the shape of a cone whose extremeties subtend an angle ϕ' with its symmetry axis, lying in the forward direction, which we take to be the x axis. In a frame in which the flashlight is moving in the positive x direction, the cone extremeties subtend an angle of ϕ. See fig. 10.2. In the frame in which the flashlight is stationary, the components of velocity of a light ray on the edge of the cone are

$$\begin{aligned}u'_x &= c\cos\phi' &(10.23)\\ u'_y &= c\sin\phi' &(10.24)\end{aligned}$$

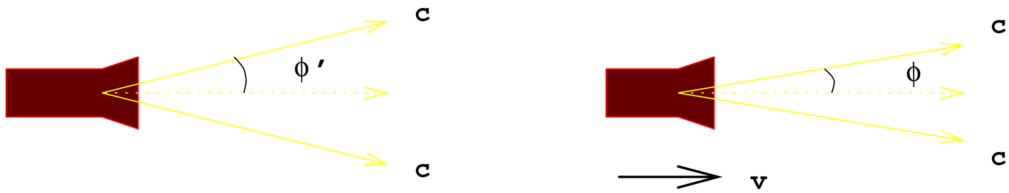

Figure 10.2: A conical light beam from a stationary source and as seen from a frame in which the source is moving.

where we have taken the direction perpendicular to the beam axis to be the y one. Similarly for the velocity components of that same ray in the frame in which the flashlight moves with speed v:

$$u_x = c\cos\phi \tag{10.25}$$
$$u_y = c\sin\phi. \tag{10.26}$$

These components are related by the velocity transformation laws, Eqs. 10.10 and 10.11, whose inverses are

$$u_x = \frac{u'_x + v}{1 + \frac{vu'_x}{c^2}} \tag{10.27}$$

and

$$u_y = \frac{u'_y}{\gamma(1 + \frac{vu'_x}{c^2})}. \tag{10.28}$$

The inverse transformation used here can be seen to be relevant by noting that the velocity of the source should increase the x component of velocity and thus a + sign is expected in the numerator of Eq. 10.27.

Substituting the values of the velocity components in terms of the sines and cosines of the angles, and dividing by c, we obtain

$$\cos\phi = \frac{\cos\phi' + \beta}{1 + \beta\cos\phi'} \tag{10.29}$$

and

$$\sin\phi = \frac{\sin\phi'}{\gamma(1 + \beta\cos\phi')}. \tag{10.30}$$

It is easy to see from the last equation that for $\phi' < 90°$ we must have $\phi < \phi'$. In fact $\phi < \phi'$ even for ϕ' up to 180°. Thus the beam is concentrated in the forward direction.

One might naively have expected the cone to broaden rather than narrow (and to do so symmetrically about 90°), in analogy with the tilted rod of fig. 5.5. However, the two situations are not analogous. With the rod, the coordinates of both ends are taken at the same time and so the phenomenon of length contraction in the direction of motion is applicable. However, a light ray moves at speed c and there is a time separation between different points on the ray.

Exercise 10.2 *Let the cone in the above originate from the origin and consider another point (x', y') on the edge of the cone. Show that if a light ray starts from the origin at $t' = 0$ then it reaches (x', y') at $t' = x'/(c\cos\phi') = y'/(c\sin\phi')$. Now use the Lorentz transformations to determine the corresponding points in the frame in which the flashlight is moving. Hence show that*

$$\tan\phi = \frac{y}{x} = \frac{\tan\phi'}{\gamma(1+\beta/\cos\phi')} \tag{10.31}$$

and verify that this is consistent with the above results.

The headlight effect is in fact a quite commonly observed phenomenon in high energy particle physics. Suppose that an object radiates uniformly in every direction while at rest. Then half of the radiation will go one way and half the other. The half that goes in the forward direction corresponds to $\phi' = 90°$ in the above. In this case $\sin\phi' = 1$, $\cos\phi' = 0$ and we obtain $\sin\phi = 1/\gamma$. Thus, as v increases to a substantial fraction of c, the time dilation factor γ increases and this half of the radiation becomes concentrated in a narrow cone in the forward direction. For example, if $v = 0.99\,c$ then $\gamma = 7.1$ and so $\phi \approx 8.1°$. Indeed, even radiation that is emitted in the negative x' direction in the source's rest frame tends to get sent in the forward direction in the moving frame. This concentrating of emitted radiation is seen in the synchrotron radiation of accelerated charged particles in high energy particle accelerators. An analogous effect is also seen in the decay products of particles created in particle accelerators. When the decaying particle has a high velocity, most of the particles appearing amongst the decay products are concentrated in the forward direction.

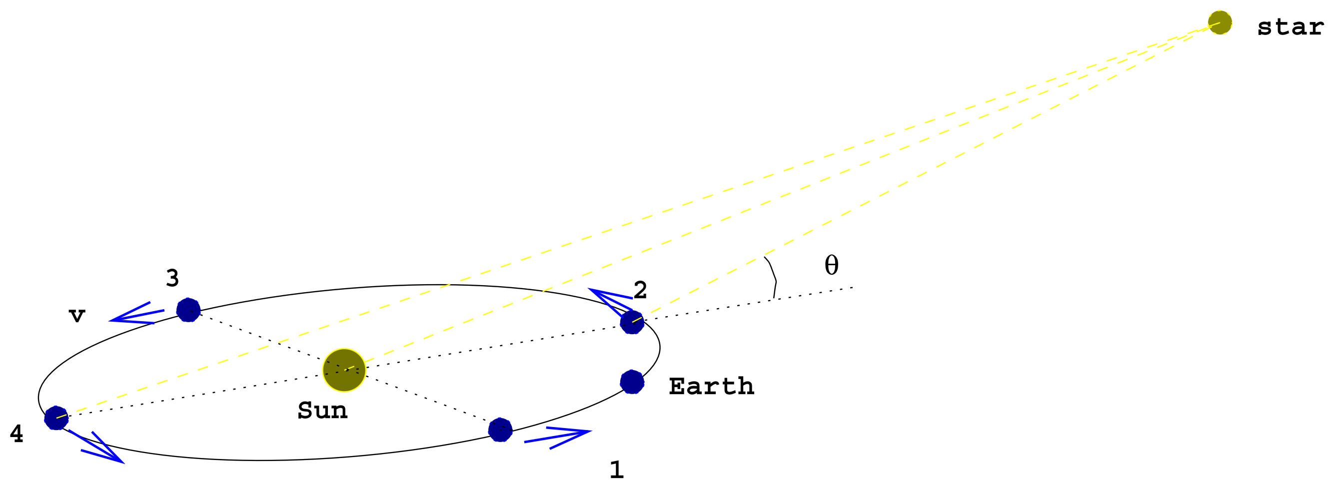

Figure 10.3: The changing elevation, θ, of a star above Earth's orbital plane as Earth moves in its orbit. (Not to scale.)

10.4 Stellar aberration

Stellar aberration was first observed by the British astronomer James Bradley in 1725. Bradley was looking for parallax effects. Parallax refers to the apparent motion of nearby objects against the background of distant objects as the observer moves. (Hold a pen at arms length and look at the projection of the pen against a far wall as you move your head. The changing projection is parallax.) Knowing the amount of parallax, by measuring the angular changes with respect to the distant background, and knowing the diameter of Earth's orbit, one has sufficient information to enable a simple calculation of the distance to the nearest stars.

The parallax effect can be understood from fig. 10.3. The maximum elevation above Earth's orbital plane should occur at position 2 and the minimum elevation at position 4, with positions 1 and 3 yielding different lateral angles.

However, Bradley found the maximum elevation occurred at position 3 and the minimum occured at position 1. The observed effect was of the order of $40''$ of arc. It turns out that the nearest stars are so far away that their parallax is less than $1''$. Although Bradley didn't know the size of parallax effects, the discrepancy in the positions at which maximum and minimum

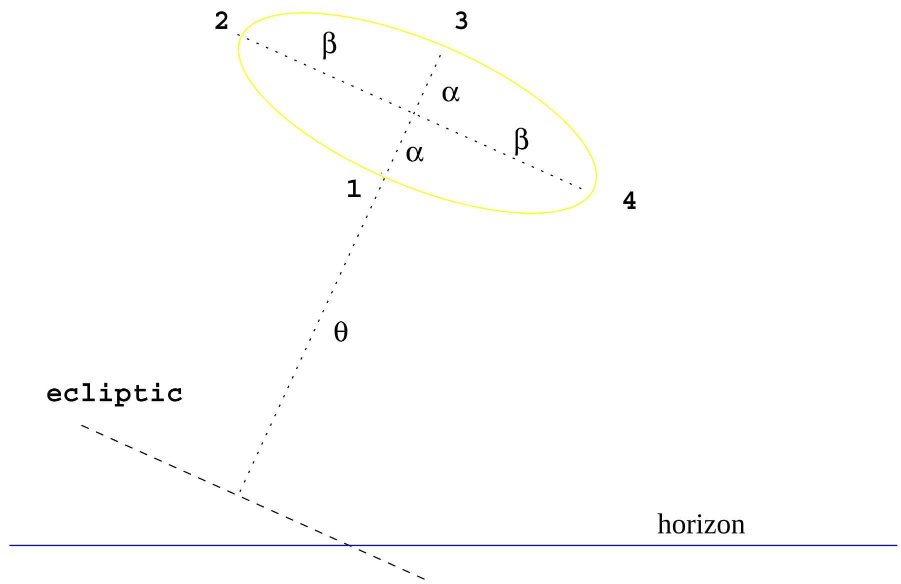

Figure 10.4: The changing ecliptic coordinates of a star over the course of a year. (Not to scale.)

elevation occurred indicated that a new phenomenon had been discovered. We call it stellar aberration. It is also observable for all stars, not just a few close ones.

The elevation, θ, of the star above the plane of Earth's orbit is known as the ecliptic latitude. The ecliptic is the name given to the plane of Earth's orbit and it's projection on the background stars. This is very well determined. Thus a star may be observed to describe a small ellipse in the sky, over the course of a year, if its ecliptic coordinates can be accurately measured. The semi-minor axis of this ellipse is usually called α and the semi-major axis β. See fig. 10.4. It is the effect of α that leads to the variation in ecliptic latitude.

The explanation of stellar aberration follows from the velocity transformations. Suppose a star is observed via a telescope at rest. Then the component of velocity, of light from the star, in the direction of the plane of the ecliptic is

$$u'_x = -c \cos\theta' \tag{10.32}$$

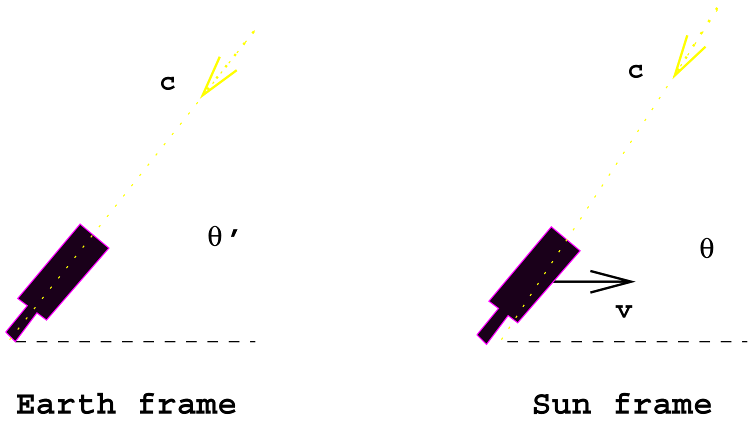

Figure 10.5: Changing elevation (above the ecliptic) of a star as judged by an observer using a moving telescope, compared to a stationary one.

while the component of velocity perpendicular to the ecliptic is

$$u'_y = -c\,\sin\theta'. \tag{10.33}$$

In the frame of the Sun, in which the telescope is moving with speed v the velocity components are

$$u_x = -c\,\cos\theta \tag{10.34}$$
$$u_y = -c\,\sin\theta. \tag{10.35}$$

At position 1 in Earth's orbit, the telescope is moving towards the star with speed v in Sun's frame. See fig. 10.5. Note that in this frame, the moving telescope must point in a slightly different direction than the incident light in order for the light to pass down its axis to the eyepiece as the telescope moves past. We take the direction of v to be along the positive x axis and the y direction is that perpendicular to it. In this case the components of velocity of the light from the star are related by the velocity transformations, Eqs. 10.10 and 10.11. Since the person looking through the telescope is at rest with respect to it, these are appropriate. Thus

$$u'_x = \frac{-c\cos\theta - v}{1+\beta\cos\theta} \tag{10.36}$$
$$u'_y = \frac{-c\sin\theta}{\gamma(1+\beta\cos\theta)}. \tag{10.37}$$

Exercise 10.3 *Verify that ${u'_x}^2 + {u'_y}^2 = c^2$ as required by the constancy of c.*

As we did in our discussion of the headlight effect, we now substitute for u'_x and u'_y in terms of $\cos\theta'$ and $\sin\theta'$ to obtain

$$\cos\theta' = \frac{\cos\theta + \beta}{1 + \beta\cos\theta} \tag{10.38}$$

and

$$\sin\theta' = \frac{\sin\theta}{\gamma(1 + \beta\cos\theta)}. \tag{10.39}$$

The difference here with the results for the headlight effect is that in this case the light is headed in the opposite direction (relative to the direction of v).

Exercise 10.4 *Interpret stellar aberration in terms of the time, Δt, taken for light to enter the top of the telescope and arrive at the eyepiece at the bottom. In the frame in which the telescope is moving, the eyepiece will have moved $v\Delta t$ whereas the light will have travelled $c\,\cos\theta\,\Delta t$ towards the eyepiece in the x direction and $c\,\sin\theta\,\Delta t$ in the y direction. (Here, θ is the angle of elevation of the star as seen by a stationary telescope in that same frame.) Show that the telescope must subtend the angle*

$$\theta_v = \tan^{-1}\frac{c\,\sin\theta}{v + c\,\cos\theta}$$

and use the analogy with the tilted stick discussed in chapter 5 to show that this is consistent with Eqs. 10.38-10.39.

For Earth in its orbit, $v = 30\text{km/s}$ and so $\beta = 10^{-4}$. Since $\beta \ll 1$ we may approximate (using the binomial theorem on Eq. 10.39 and neglecting terms small compared to β)

$$\sin\theta' \approx \sin\theta(1 - \beta\cos\theta) \tag{10.40}$$

Now let

$$\theta' = \theta - \alpha. \tag{10.41}$$

Then it is a general result of trigonometry that

$$\sin\theta' = \sin\theta\cos\alpha - \cos\theta\sin\alpha \tag{10.42}$$

$$\approx \sin\theta - \alpha\cos\theta \text{ for } \alpha \text{ small.} \tag{10.43}$$

Thus, by comparison, the angle of aberration is

$$\alpha \approx \beta\sin\theta. \tag{10.44}$$

At position 3 the sign of β reverses. These results are thus in accord with the observation that the ecliptic latitude of the star is a minimum at 1 and a maximum at 3.

At position 4, take the velocity of Earth to be in the positive z direction in Sun's frame. As before, the components of starlight velocity in Sun's frame are $u_x = -c\cos\theta$, $u_y = -c\sin\theta$ and $u_z = 0$. In the telescope frame:

$$u'_x = \frac{u_x}{\gamma(1 - \frac{vu_z}{c^2})} = \frac{-c\cos\theta}{\gamma} \tag{10.45}$$

$$u'_y = \frac{u_y}{\gamma(1 - \frac{vu_z}{c^2})} = \frac{-c\sin\theta}{\gamma} \tag{10.46}$$

$$u'_z = \frac{u_z - v}{(1 - \frac{vu_z}{c^2})} = -v. \tag{10.47}$$

Exercise 10.5 *Verify that ${u'_x}^2+{u'_y}^2+{u'_z}^2 = c^2$. Show also that the component of velocity in the xy plane is*

$$u'_\perp = c/\gamma \tag{10.48}$$

If we call the deviation of the direction of arrival of the starlight away from the xy plane β then

$$\tan\beta = u'_z/u'_\perp = -\frac{v}{c}\gamma \approx -\frac{v}{c}. \tag{10.49}$$

For β small, we thus have

$$\tan\beta \approx \beta \approx -\frac{v}{c}. \tag{10.50}$$

At position 2 we similarly find

$$\beta \approx +\frac{v}{c}. \tag{10.51}$$

(That β is conventionally used for this ratio of velocities is no accident but stems from the early study of stellar aberration.) These results, with their signs are thus in accord with the situation depicted in fig. 10.4.

10.4.1 Airy's experiment

In 1871 Sir George Airy conducted an experiment in which stellar aberration was observed through a water filled telescope. He found that the amount of aberration was exactly the same as with an ordinary telescope. This null result was quite puzzling in the old ether theory and had to be explained by an extraordinary "partial drag of the ether" by the water, in which Fresnel's drag coefficient (see section 10.2) arose for inexplicable reasons.

However, the null result is readily understood in Einstein's theory in which the speed of light is independent of the velocity of the observer. The easiest way to grasp the null result is to consider the frame in which the telescope is stationary (see fig. 10.5.) In this frame, the starlight passes directly down the telescope axis without any refraction by the water. Thus the aberration angle is given directly by the velocity transformations for the incident light on the telescope objective, independently of whatever medium fills the telescope interior.

10.5 The 4-velocity

The complex nature of the velocity transformations stems from the transformation of the time variable in the denominator. However, there are other ways of describing the rate of an object's motion. In chapter 2 we saw that a traveller in a car could measure the speed at which the road passed by via the formula $v_{\text{road}} = \Delta s/\Delta\tau$ where s is the length of road that passes by and τ is the proper time recorded, for example, on the traveller's wristwatch. The proper time, though only applicable in one frame, is something that all observers can calculate and agree on as it is essentially just the spacetime interval, which is Lorentz invariant. Thus, if we define the 4-velocity as

$$u^\mu = \frac{dx^\mu}{d\tau} \qquad \mu = 0, 1, 2, 3, \tag{10.52}$$

where $(x^0, x^1, x^2, x^3) = (ct, x, y, z)$, then in another frame moving in the positive x direction with respect to the first the 4-velocity will be

$$\begin{aligned} u'_x = \frac{dx'}{d\tau} &= \frac{\gamma(dx - \beta cdt)}{d\tau} \\ &= \gamma(\frac{dx}{d\tau} - \beta c\frac{dt}{d\tau}) \\ &= \gamma(u_x - \beta u^0) \end{aligned} \tag{10.53}$$

$$\begin{aligned} u'_y = \frac{dy'}{d\tau} &= \frac{dy}{d\tau} \\ &= u_y \end{aligned} \tag{10.54}$$

$$\begin{aligned} u'_z = \frac{dz'}{d\tau} &= \frac{dz}{d\tau} \\ &= u_z \end{aligned} \tag{10.55}$$

$$\begin{aligned} u'^0 = \frac{dx'^0}{d\tau} &= \frac{\gamma(cdt - \beta dx)}{d\tau} \\ &= \gamma(\frac{cdt}{d\tau} - \beta\frac{dx}{d\tau}) \\ &= \gamma(u^0 - \beta u_x). \end{aligned} \tag{10.56}$$

It is to be noted that the components of the 4-velocity transform in exactly the same way as the position vector. This makes it particularly convenient to deal with.

The zero (i.e. time) component of the 4-velocity may be evaluated by noting that t is simply related to τ by the time dilation formula 6.3. Thus

$$dt = \gamma_u d\tau$$

and

$$u^0 = \frac{d(ct)}{d\tau} = \gamma_u c \tag{10.57}$$

where γ_u is the time dilation factor for an object moving with speed u and is to be distinguished from the γ factor in the transformation equations above, which involves the relative speed v of the two frames.

Exercise 10.6 *Show*

$$(u^0, \mathbf{u}) = \gamma_u(c, \mathbf{v}) \tag{10.58}$$

Review

Derive the velocity transformations from one frame to another moving relative to the first with speed v along the positive z direction.

How is the velocity of light in a medium related to that in vacuum? Why are they different?

What is Cerenkov radiation?

In the Fizeau experiment light is sent around a path in flowing water in two different directions so that in one case the light travels upstream and in the other it travels downstream. An interference pattern is observed when the two light beams recombine. Compared to the case where the water is stationary rather than flowing, is a fringe shift observed or not?

What is the headlight effect?

What is *stellar aberration*?

What is the *ecliptic*?

What was Airy's experiment and its result?

The relativistic velocity transformations are much more complicated than the Galilean velocity addition rule of Newtonian physics. Explain why this is so.

What is the 4-velocity?

Questions

1. Do the observations of Arago, on the refraction of starlight, and Airy, on stellar aberration, provide good evidence for the independence of c on source velocity?

2. What does the zero component of the 4-velocity represent?

Problems

1. In the 25th century a space patrol is in hot pursuit of a spaceship fleeing from a spacestation prison. As measured by an observer on the spacestation the speeding spaceship is travelling at $0.4c$. and the pursuing patrol is chasing at $0.5c$.

   (a) What is the speed of the fleeing spaceship as measured by the patrol ship?

   (b) The patrol ship switches on its searchlight. At what speed do observers on the spacestation see the beam travel?

2. It is found experimentally that if light is shone on a mirror at rest then the angle of incidence, θ_i (defined as the angle between the incident beam and a perpendicular to the mirror), and the angle of reflection, θ_r (defined as the angle between the reflected beam and the normal to the mirror), are equal (and that the incident and reflected beams lie in a plane). Consider the arrangement shown in the figure, where a mirror is inclined at 45° to the x and y axes. A light beam shining down the y-axis is thus reflected along the x-axis. In a frame in which the mirror is moving along the positive x-axis with speed v

   (a) what are the directions of the incident and reflected beams?

   (b) What is the orientation of the mirror?

   (c) Has θ_r increased, decreased or stayed the same?

   (d) Has θ_i increased, decreased or stayed the same?

   (e) Is $\theta_r = \theta_i$ in the new frame?

3. Experimentally, it is found that light entering a stationary medium is bent towards the normal to the surface (see fig. 10.7). The phenomenon is known as refraction and the amount of bending is given by Snell's formula

$$\frac{\sin\theta'_i}{\sin\theta'_r} = n \tag{10.59}$$

   where n is the index of refraction. (Note that the incident and refracted rays lie in a plane.) Consider refraction in the case where the medium is moving with respect to the observer.

(a) Let the medium move with speed v in a direction parallel to its surface and in the plane containing the light ray.

   i. Use velocity transformations to relate the components of velocity of both the incident and refracted light to those in a frame in which the medium is at rest.

   ii. What is the speed of light in the moving medium?

   iii. Relate the angles of incidence and refraction. Is Snell's formula applicable?

(b) Repeat for the case where the motion is perpendicular to the plane of the light ray.

(Note that Arago's observation, mentioned in the text, that starlight is refracted by glass as if the glass is not moving corresponds to the fact that indeed the glass was not moving with respect to the observer and that the speed of light in vacuum is the same for all inertial observers. Nevertheless, the relativistic calculations for the moving medium go a long way to explaining why the mysterious, and incorrect, "partial drag" calculations of the old ether theory worked.)

4. Analyze Airy's experiment in the frame in which the telescope is moving. Note that in this frame the telescope must be tilted away from the direction of incident starlight in order for the light to arrive at the eyepiece. Thus refraction at the telescope objective occurs. Show that when the velocity of light in moving water and refraction are properly accounted for, the same angle of tilt is required as in the case that the telescope has no water inside.

5. Consider the twin paradox. In order to avoid complications with accelerating clocks, let there be two travellers: an outward bound one travelling at β, and an inward bound one travelling at $-\beta$, who pass each other at the "return point", located a distance D from the stay-at-home twin. (All of these values pertain to the stay-at-home twin's frame.) In order to determine the elapsed time for the stay-at-home twin, the travellers must not only measure their journey times but a communication of information must take place between them. This information must consist of both the time at which they pass and (for the requirements of the homeward bound traveller) the time at which the outward bound traveller left home.

For convenience, let the origin of both travellers' frames be at the return point.

- Show that in the outward bound traveller's frame, home is located at $x = 0$ and he left there at

$$ct = -\frac{D}{\gamma\beta}.$$

- What is the velocity, β', of the homeward bound traveller in the outward bound traveller's frame?
- What is the location of home in the homeward bound traveller's frame? In this frame, at what time did the outward bound traveller leave home? Show that in this same frame that the outward bound journey took

$$ct'' = \frac{\gamma(1+\beta^2)D}{\beta}$$

  and that the combined time for the outward bound and homeward bound trips is

$$ct'' = \frac{\gamma 2D}{\beta}.$$

  Infer the corresponding total elapsed time at home before the homeward bound traveller gets there, i.e. the time measured by the stay-at-home twin, and show that this is exactly as one would expect for the inertial stay-at-home observer.

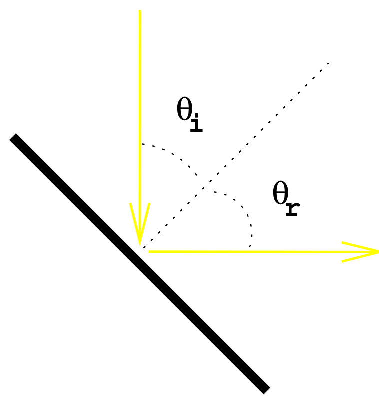

Figure 10.6: Reflection arrangement.

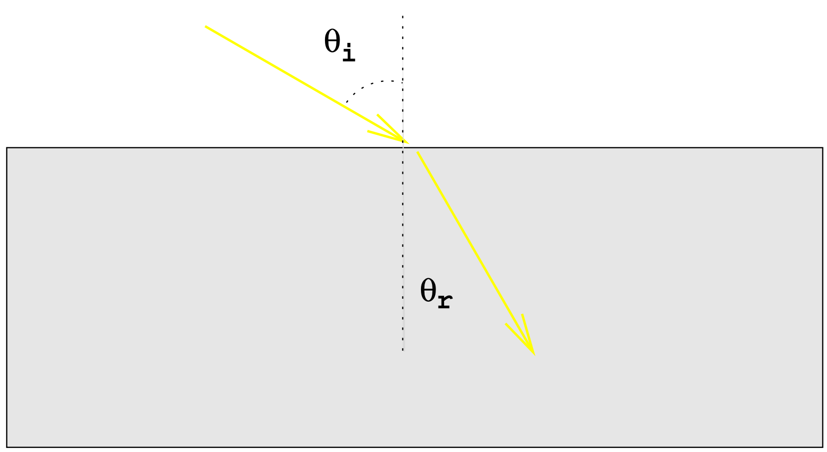

Figure 10.7: The refraction of light on entering a medium

Part II

Dynamics

Chapter 11

Invariance in Particle Collisions

We now consider what happens when two or more particles collide with each other. For simplicity, we assume the particles to be pointlike and to obey the classical particle axiom. We exclude from the present consideration, any interaction between the particles, except the contact interaction that takes place at the instant of collision. We also assume that all particles are free from influences that are external to the system of particles, i.e. the only interactions between particles are those that arise in (contact) collisions of the particles with each other.

The mechanical details of the interaction taking place in a contact collision are accessible only through analysis of collision experiments. Those details are not our present concern though; indeed a contact collision is somewhat of an idealization. Rather, our aim here is to determine what overall restrictions the Invariance Postulate places on the collision process. We saw in chapter 3 that translation invariance implied that Newton's First Law must hold for a single isolated particle. We investigate now whether there is an analogous result for an isolated system of particles.

11.1 Two particle collisions

Both before and after a contact collision, each particle involved obeys Newton's First Law and so moves in a straight line. During the contact collision though, the particles influence each other and so then Newton's First Law does not apply. Consequently the particle trajectories may be altered during the collision.

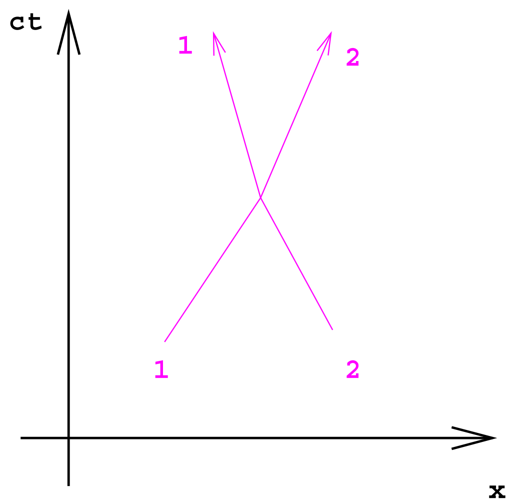

Figure 11.1: Spacetime diagram for a two-particle contact collision showing the x component of the particle worldlines as time evolves.

We seek to characterize what happens during the collision. The only information we have is that provided by the worldlines of the particles. Thus we know the positions of both particles at any instant and we know their velocities before and after. The details of the collision process are therefore to be considered as determined by this information. See fig. 11.1.

The Invariance Postulate tells us that what happens at the instant of the collision does not depend on where this collision takes place — the postulate being applicable because the process is free of external influence and so an experiment of this type could be conducted in an enclosed laboratory.

In order to analyze this situation better, it is useful to understand more about how the particle coordinates are described. In particular we need to understand what other ways there may be to represent the particle coordinates and whether any others may be more suitable than the coordinates of the individual particles.

11.2 Linear independence of vectors

Consider an ordinary vector in a two dimensional Euclidean space, as in fig. 11.2. It is common to represent this vector via its coordinates:

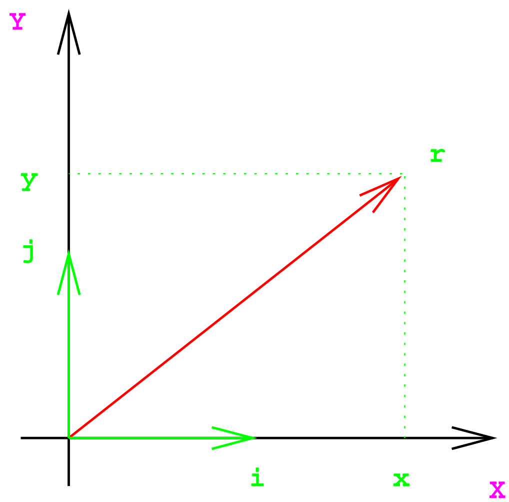

Figure 11.2: Representation of a vector in a plane in terms of basis vectors.

$$\mathbf{r} = (x, y) \equiv x\mathbf{i} + y\mathbf{j} \equiv x\mathbf{e}_x + y\mathbf{e}_y. \tag{11.1}$$

Here, $\mathbf{i} \equiv \mathbf{e}_x$ and $\mathbf{j} \equiv \mathbf{e}_y$ are unit vectors (i.e. vectors of unit length) directed along the x and y axes respectively. Thus $\mathbf{r}$ is said to be a linear combination of $\mathbf{i}$ and $\mathbf{j}$. The coefficients x and y in the combination are the coordinates. Any vector lying in the xy-plane can be so represented and the vectors $\mathbf{i}$ and $\mathbf{j}$ are said to *span* the two-dimensional space of the plane and to form a *basis* for it.

These basis vectors are by no means unique though. We saw in chapter 8 that a rotation of the coordinate axes led to new coordinates. These result because one now uses new basis vectors $\mathbf{i}'$ and $\mathbf{j}'$ obtained by rotating the old basis vectors. However, the basis vectors need not even be of unit length, or at right angles to each other. We could for instance write

$$\mathbf{r} = \frac{x}{2}\mathbf{b}_1 + 3y\mathbf{b}_2$$

where $\mathbf{b}_1 = 2\mathbf{e}_x$ and $\mathbf{b}_2 = \mathbf{e}_y/3$; or we could use $\mathbf{e}_x$ and $\mathbf{b}_3 = \mathbf{e}_x + \mathbf{e}_y$ as basis vectors.

Exercise 11.1 *Show that*

$$\mathbf{r} = (x - y)\mathbf{e}_x + y\mathbf{b}_3.$$

We could not though express a general vector $\mathbf{r}$ in terms of $\mathbf{e}_x$ and $\mathbf{b}_1 = 2\mathbf{e}_x$ because these vectors are just multiples of each other and point in the same direction. They do not therefore span the space of the xy-plane and are said to be linearly dependent.

Similarly, if we go to three space dimensions we can write a general vector as

$$\mathbf{r} = (x, y, z) \equiv x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \equiv x\mathbf{e}_x + y\mathbf{e}_y + z\mathbf{e}_z. \tag{11.2}$$

Again, $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are not unique basis vectors. We could just as well use $\mathbf{b}_1 = 2\mathbf{e}_x$, $\mathbf{b}_2 = \mathbf{e}_x + \mathbf{e}_y$ and $\mathbf{b}_3 = \mathbf{e}_y + \mathbf{e}_z$. We could not though use $\mathbf{e}_x$, $\mathbf{e}_y$ and $\mathbf{b}_2$ since the latter is just a combination of the first two and not therefore independent of them. These three vectors all lie in the xy-plane and the full space (including the z dimension) is therefore not spanned.

If one vector can be expressed as a linear combination of others then clearly we have a linear combination of all three that equals zero, i.e.

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = 0. \tag{11.3}$$

We take this as a formal definition of linear dependence of the vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$. If no non-zero coefficients exist that would make this happen then the vectors are said to be *linearly independent*. This definition can clearly be extended to other dimensions and the dimension is just the maximum number of linearly independent vectors.

11.3 Invariance in two particle collisions

Let us return now to our two-particle collision. If we consider the two particles at a given time, t, then, except at the point of collision, the spacetime separation of the particles is spacelike. Thus there can be no instantaneous communication between them and the spatial coordinates at time t must be independent of each other. We can therefore say that the spatial coordinates $\mathbf{x}_1(t) = (x_1(t), y_1(t), z_1(t))$ and $\mathbf{x}_2(t) = (x_2(t), y_2(t), z_2(t))$ are linearly independent vectors in a $3 \times 2 = 6$ dimensional *configuration space*. (For N particles we would have a $3N$ dimensional configuration space. This space may be regarded conveniently as N replications of ordinary space.)

No concept of vectors being at right angles to each other in this configuration space need be introduced. All that matters is its dimension. As seen above, the dimension of the space is just the number of basis vectors needed to span it. An obvious choice for these basis vectors are the vector components $x_1(t)\mathbf{e}_x$, $y_1(t)\mathbf{e}_y$, $z_1(t)\mathbf{e}_z$, $x_2(t)\mathbf{e}_x$, $y_2(t)\mathbf{e}_y$ and $z_2(t)\mathbf{e}_z$. This is not the only choice though and it turns out not to be the most convenient for our purposes.

11.3.1 Translation invariance

Consider the two-dimensional subspace of configuration space corresponding to the x coordinates of the two particles. It is spanned by $x_1\mathbf{e}_x$ and $x_2\mathbf{e}_x$, or any other two linearly independent combinations of them. It is also independent of the y and z subspaces.

Consider a (passive) translation by X in the x direction. Then

$$\begin{aligned} x_1 &\to x_1 - X \\ x_2 &\to x_2 - X. \end{aligned} \tag{11.4}$$

If we consider the linear combination

$$x_- \equiv x_1 - x_2 \tag{11.5}$$

then x_- is completely independent of X. However, any other linearly independent combination of x_1 and x_2 will depend on X. What this means is that x_- is an internal coordinate for the system which describes the relative separation of the particles. The other linearly independent basis vector will contain information on the location of the system as a whole.

If we consider the velocity corresponding to any vector in the two-dimensional configuration subspace, i.e. the time derivative of that vector, then that velocity is invariant under the (passive) translation by X. Hence any variation of that velocity that conveyed information on X could not be attributed to choice of the coordinate system and must mean that translation invariance is violated. How can we avoid conveying information on X when the velocities of both particles are changing? Part of the answer lies in the fact that x_- is independent of X. Hence

$$v_-^x = \frac{dx_-}{dt} \tag{11.6}$$

may depend on x_-, without being inconsistent with translation invariance. Every other linearly independent vector in the subspace depends on X. Suppose though that there exists one of them, call it $x_c\mathbf{e}_x$, for which

$$v_c^x = \frac{dx_c}{dt} \tag{11.7}$$

is independent of x_c, and hence X. Then every other vector in the 2-dimensional subspace can be written as a linear combination of $x_c\mathbf{e}_x$ and $x_-\mathbf{e}_x$ and any variation in the velocity for that vector, which is a combination of v_c^x and v_-^x would be due to the variation in v_-^x. If a vector behaving like x_c did not exist then the observed variations of the particle velocities during a collision would convey information about the location of the collision in space.

The Invariance Postulate therefore requires the existence of a vector

$$x_c\mathbf{e}_x = (c_1^x x_1 + c_2^x x_2)/N_x\,\mathbf{e}_x \tag{11.8}$$

which changes (only) at a constant rate throughout the collision. The coefficients c_1^x and c_2^x are unknown, a priori, except that they are not in the ratio $1 : -1$ and are unique to within normalization. To highlight the arbitrariness in normalization we have explicitly included a normalization factor in Eq. 11.8.

Similar considerations for the two-dimensional y and z subspaces lead us to the conclusion that there must exist vectors

$$y_c\mathbf{e}_y = (c_1^y y_1 + c_2^y y_2)/N_y\,\mathbf{e}_y \tag{11.9}$$
$$z_c\mathbf{e}_z = (c_1^z z_1 + c_2^z z_2)/N_z\,\mathbf{e}_z \tag{11.10}$$

that change at constant rates before, during and after the collision. In other words, the velocities

$$v_c^y = \frac{dy_c}{dt} \tag{11.11}$$
$$v_c^z = \frac{dz_c}{dt} \tag{11.12}$$

must remain constant throughout the collision. (Note that altering the normalizations of x_c, y_c and z_c will not affect the constancy of the corresponding velocities.)

Consider now the vector $\mathbf{x}_c = (x_c, y_c, z_c)$. It should be clear that this vector moves uniformly in a straight line before, during and after the collision, since its displacement in equal time intervals, dt, is always the constant vector, $d\mathbf{x}_c = (v_c^x, v_c^y, v_c^z)dt$. However, we observe that because of the arbitrary norms of the components, $\mathbf{x}_c$ is basically arbitrary, including in the direction in which it points! To bring some definiteness to the picture, let us stipulate that

$$\begin{aligned}
\frac{c_1^x}{N_x} + \frac{c_2^x}{N_x} &= 1 \\
\frac{c_1^y}{N_y} + \frac{c_2^y}{N_y} &= 1 \\
\frac{c_1^z}{N_z} + \frac{c_2^z}{N_z} &= 1. \qquad (11.13)
\end{aligned}$$

Exercise 11.2 *Show that this normalization convention is the only one for which* $\mathbf{x}_c$ *passes through the point of collision.*

Henceforth we shall take $\mathbf{x}_c$ to be this unique vector.

11.3.2 Effect of rotations

In addition to translation invariance, the collision process must be rotationally invariant. If we consider the vector $\mathbf{x}_c$ in ordinary space, then it gets altered by a rotation of the coordinate system. It is not therefore a quantity that in itself conveys any information on intrinsic orientation in space. However, the translation invariance must be consistent with rotational symmetry.

We have seen that the linear combinations forming the components of $\mathbf{x}_c$ are unique. Having selected these components, the corresponding velocity $\mathbf{v}_c$ must still remain constant after a rotation of the coordinate system. Since the vector exhibiting translation invariance of the rotated system is also unique it must be just the rotated $\mathbf{x}_c$. Furthermore, the linear combination must be independent of the orientation. If this constraint were not respected then it would imply something distinguishable about different orientations.

Consider a rotation in the xy plane. Then the rotated components of $\mathbf{x}_c$ are

$$x'_c = \cos\theta\, x_c + \sin\theta\, y_c \qquad (11.14)$$

$$y'_c = -\sin\theta\, x_c + \cos\theta\, y_c. \qquad (11.15)$$

These are to be compared with

$$
\begin{aligned}
{x'}_c &= (c_1^x {x'}_1 + c_2^x {x'}_2)/N_x && (11.16)\\
{y'}_c &= (c_1^y {y'}_1 + c_2^y {y'}_2)/N_y && (11.17)
\end{aligned}
$$

in which the coefficients must be the same as in the unrotated system, as explained above. Equating these, one finds

$$
\begin{aligned}
&\cos\theta\Big[\frac{c_1^x}{N_x}x_1 + \frac{c_2^x}{N_x}x_2\Big] + \sin\theta\Big[\frac{c_1^y}{N_y}y_1 + \frac{c_2^y}{N_y}y_2\Big]\\
&\qquad = \cos\theta\Big[\frac{c_1^x}{N_x}x_1 + \frac{c_2^x}{N_x}x_2\Big] + \sin\theta\Big[\frac{c_1^x}{N_x}y_1 + \frac{c_2^x}{N_x}y_2\Big] && (11.18)\\
&-\sin\theta\Big[\frac{c_1^x}{N_x}x_1 + \frac{c_2^x}{N_x}x_2\Big] + \cos\theta\Big[\frac{c_1^y}{N_y}y_1 + \frac{c_2^y}{N_y}y_2\Big]\\
&\qquad = -\sin\theta\Big[\frac{c_1^y}{N_y}x_1 + \frac{c_2^y}{N_y}x_2\Big] + \cos\theta\Big[\frac{c_1^y}{N_y}y_1 + \frac{c_2^y}{N_y}y_2\Big] && (11.19)
\end{aligned}
$$

and thus we must have

$$
\begin{aligned}
\frac{c_1^y}{N_y}y_1 + \frac{c_2^y}{N_y}y_2 &= \frac{c_1^x}{N_x}y_1 + \frac{c_2^x}{N_x}y_2 && (11.20)\\
\frac{c_1^x}{N_x}x_1 + \frac{c_2^x}{N_x}x_2 &= \frac{c_1^y}{N_y}x_1 + \frac{c_2^y}{N_y}x_2. && (11.21)
\end{aligned}
$$

Rearranging,

$$
\begin{aligned}
\Big(\frac{c_1^y}{N_y} - \frac{c_1^x}{N_x}\Big)y_1 &= \Big(\frac{c_2^x}{N_x} - \frac{c_2^y}{N_y}\Big)y_2 && (11.22)\\
\Big(\frac{c_1^x}{N_x} - \frac{c_1^y}{N_y}\Big)x_1 &= \Big(\frac{c_2^y}{N_y} - \frac{c_2^x}{N_x}\Big)x_2. && (11.23)
\end{aligned}
$$

Since x_1 is independent of x_2 and y_1 is independent of y_2 we must have that their coefficients in these equations are zero. Thus

$$
\begin{aligned}
\frac{c_1^x}{N_x} &= \frac{c_1^y}{N_y} && (11.24)\\
\frac{c_2^x}{N_x} &= \frac{c_2^y}{N_y} && (11.25)
\end{aligned}
$$

or,

$$\frac{c_1^x}{c_1^y} = \frac{c_2^x}{c_2^y} = \frac{N_x}{N_y}. \tag{11.26}$$

Similar considerations for rotations in the yz or zx planes lead us to conclude that consistency demands

$$\frac{c_1^x}{c_2^x} = \frac{c_1^y}{c_2^y} = \frac{c_1^z}{c_2^z} \tag{11.27}$$

as well as

$$\frac{c_1^x}{N_x} = \frac{c_1^y}{N_y} = \frac{c_1^z}{N_z} \equiv \frac{c_1}{N} \tag{11.28}$$

and

$$\frac{c_2^x}{N_x} = \frac{c_2^y}{N_y} = \frac{c_2^z}{N_z} \equiv \frac{c_2}{N}. \tag{11.29}$$

Thus consistency with rotational symmetry requires that the weighting coefficients are the same for each component of $\mathbf{x}_c$. Note that while the weighting coefficients c_i/N are unique, the scale and units for the c_i have not been fixed, nor need they be for the purposes of the present discussion.

Hence we have shown that the linear combination

$$\mathbf{x}_c(t) = (c_1\mathbf{x}_1(t) + c_2\mathbf{x}_2(t))/N \tag{11.30}$$

retains its form under rotations and the unique weights c_1/N and c_2/N are common to each component and invariant under rotations. For $N = c_1 + c_2$, this vector is the only one of this type that both passes through the collision point and for which

$$\mathbf{v}_c = \frac{d\mathbf{x}_c}{dt} = (c_1\mathbf{v}_1 + c_2\mathbf{v}_2)/N \tag{11.31}$$

is constant throughout the collision. The existence of $\mathbf{x}_c$ is a consequence of translation invariance, but the constraints on its form come from including rotational symmetry (and stipulating that it pass through the collision point). For reasons that shall become apparent in the next chapter, $\mathbf{x}_c$ is sometimes called the *center of energy*.

11.3.3 The time components and effect of boosts

Suppose that we had used variations with respect to proper time in the above. The proper velocities

$$\mathbf{u}_1 = \frac{d\mathbf{x}_1}{d\tau} \tag{11.32}$$

$$\mathbf{u}_2 = \frac{d\mathbf{x}_2}{d\tau} \tag{11.33}$$

are also independent of a translation of origin. Therefore, similar considerations to those above lead us to the conclusion that there must exist a unique system vector

$$\mathbf{x}_s(t) = (m_1\mathbf{x}_1(t) + m_2\mathbf{x}_2(t))/N_s, \tag{11.34}$$

passing through the collision point for the case $N_s = m_1 + m_2$, and where the coefficients m_1 and m_2 are rotationally (and translationally) invariant, and such that

$$\mathbf{u}_s = \frac{d\mathbf{x}_s}{d\tau} = (m_1\mathbf{u}_1 + m_2\mathbf{u}_2)/N_s \tag{11.35}$$

is constant throughout the collision. In this case though the coefficients in the linear combinations differ from those of the last section because the proper velocity is not the same as the ordinary velocity. Thus the time derivative of $\mathbf{x}_s$ will not be constant though $\mathbf{u}_s$ is.

It turns out that the vector $\mathbf{x}_s$, defined via derivatives with respect to proper time, is more elegant, as it allows us a simple comparison of the vectors in different inertial frames. While the vector $\mathbf{x}_c$, for which $\mathbf{v}_c$ is constant, exists in every inertial frame the complicated nature of the velocity transformations means that the vectors in one inertial frame are not simply related to those in another. In particular, the transformed velocity vector from one frame may not be constant in another.

It transpires though that the normalization $m_1 + m_2$ is not the most convenient as it inhibits elegant incorporation of the time component of the 4-velocity. Let us therefore explore alternative normalizations. One notes from the relationship between 4-velocity and ordinary velocity that

$$\mathbf{u}_s = (m_1\gamma_1\mathbf{v}_1 + m_2\gamma_2\mathbf{v}_2)/N_s. \tag{11.36}$$

The subscripts here on the time dilation factors γ refer to the individual particles, which have different speeds in general. We can also write

$$\mathbf{u}_s = \gamma_s \mathbf{v}_s \tag{11.37}$$

and if $\mathbf{u}_s$ is constant then so will be $\mathbf{v}_s$. If we integrate this $\mathbf{v}_s$ with respect to t rather than τ then we obtain

$$\mathbf{x}(t) = (m_1\gamma_1\mathbf{x}_1(t) + m_2\gamma_2\mathbf{x}_2(t))/(\gamma_s N_s) \tag{11.38}$$

plus an integration constant. Choosing this constant to be zero and

$$N_s = (m_1\gamma_1 + m_2\gamma_2)/\gamma_s \equiv M_s \tag{11.39}$$

results in $\mathbf{x}(t)$ passing through the collision point and by construction it does so at a uniform rate. Note that the position coordinate $\mathbf{x}$ is not the same as $\mathbf{x}_s$ because increments in τ correspond to different increments in t for the two particles. Nevertheless, $\mathbf{x}$ is uniquely obtained from $\mathbf{x}_s$ and and must just be the unique vector $\mathbf{x}_c$ discussed in the last section. The uniqueness of $\mathbf{x}_c$ and $\mathbf{x}_s$ thus requires that

$$\frac{m_1\gamma_1}{\gamma_s N_s} = \frac{c_1}{c_1+c_2} \quad \text{and} \quad \frac{m_2\gamma_2}{\gamma_s N_s} = \frac{c_2}{c_1+c_2}, \tag{11.40}$$

implying

$$\frac{m_1\gamma_1}{m_2\gamma_2} = \frac{c_1}{c_2}, \tag{11.41}$$

a possible solution to which is

$$c_1 = \gamma_1 m_1 \tag{11.42}$$

$$c_2 = \gamma_2 m_2. \tag{11.43}$$

Any constant multiplying this will also yield a suitable solution and the choice c^2 turns out to sometimes be useful. The normalization in Eq. 11.39 means that $\mathbf{x}_s$ will not pass through the collision point but will run parallel to the vector with norm $m_1 + m_2$ that does. Its main benefit is that it permits us to simply and elegantly incorporate the time components of the system 4-velocity in Eq. 11.35. Let us turn our attention now to the time components and focus on the 4-velocities.

Observe in the above that we considered the positions of each particle at the same time. Thus

$$t_1 = t_2 = t_s = t \tag{11.44}$$

and the time components of the particle worldlines are clearly not linearly independent of each other. It seems to make little sense therefore to consider linear combinations of the time coordinates. However, the rate at which each of these times changes with respect to proper time (for the respective spatial coordinates) is not the same. Thus the time components of the 4-velocities differ. In fact we can write

$$u_s^0 = \frac{d(ct_s)}{d\tau} = \gamma_s c \tag{11.45}$$

$$u_1^0 = \frac{d(ct_1)}{d\tau} = \gamma_1 c \tag{11.46}$$

$$u_2^0 = \frac{d(ct_2)}{d\tau} = \gamma_2 c \tag{11.47}$$

where we use the fact that $d\tau$ is the same for each particle. Consequently,

$$u_s^0 = \frac{\gamma_s}{\gamma_1} u_1^0 = \frac{\gamma_s}{\gamma_2} u_2^0. \tag{11.48}$$

Clearly the time components of all 4-velocities are proportional to each other. However, this does not mean that they are linearly dependent. The proportionality coefficients are actually complicated functions of *both* particles' velocities and they may even change during the collision. In fact the time components of the 4-velocities must be independent of each other for space-like separation of the particles as one particle cannot know the speed of the other.

We note that for any 4-velocity, the time component, u^0, obeys

$$(u^0)^2 - \mathbf{u}^2 = c^2. \tag{11.49}$$

Since $\mathbf{u}_s$ remains constant during the collision, so must u_s^0 — and it is uniquely determined if $\mathbf{u}_s$ is. Thus we do expect that u_s^0 is a linear combination of u_1^0 and u_2^0:

$$u_s^0 = (m_1^0 u_1^0 + m_2^0 u_2^0)/N_0. \tag{11.50}$$

Whatever the coefficients in this combination are, consistency of Eqs. 11.48 and 11.50 requires that

$$N_0 = (m_1^0\gamma_1 + m_2^0\gamma_2)/\gamma_s. \tag{11.51}$$

This compulsory constraint serves an analogous role in pinning down the normalization of the time component as the arbitrary constraint that $\mathbf{x}_c$ passes through the collision point plays in pinning down the normalization of the spatial components.

Suppose we were in another inertial frame moving uniformly with respect to the first. Then the constant 4-velocity of the system in the new frame is given by the Lorentz transformations

$$\begin{aligned} u'^0_s &= \gamma(u^0_s - \beta u^x_s) &(11.52)\\ u'^x_s &= \gamma(u^x_s - \beta u^0_s). &(11.53)\end{aligned}$$

As was the case with rotations, invariance under Lorentz boosts requires that this be consistent with

$$\begin{aligned} u'^0_s &= (m_1^0 u'^0_1 + m_2^0 u'^0_2)/N_0 &(11.54)\\ u'^x_s &= (m_1 u'^x_1 + m_2 u'^x_2)/N_s &(11.55)\end{aligned}$$

in which the coefficients are the same as for u^μ_s in the original frame.

Exercise 11.3 *Show that consistency requires*

$$\begin{aligned} \left(\frac{m_1^0}{N_0} - \frac{m_1}{N_s}\right)u_1^x &= \left(\frac{m_2}{N_s} - \frac{m_2^0}{N_0}\right)u_2^x &(11.56)\\ \left(\frac{m_1}{N_s} - \frac{m_1^0}{N_0}\right)u_1^0 &= \left(\frac{m_2^0}{N_0} - \frac{m_2}{N_s}\right)u_2^0. &(11.57)\end{aligned}$$

Since u_1^μ and u_2^μ are independent of each other we must have that their coefficients in the above equations are zero. Thus

$$\frac{m_1^0}{N_0} = \frac{m_1}{N_s} \text{ and } \frac{m_2^0}{N_0} = \frac{m_2}{N_s}. \tag{11.58}$$

The coefficients in Eq. 11.50 are therefore the same as those for the space components. Note that the normalizations of Eqs. 11.39 and 11.51 are identical if we make the Poincaré invariant selection $m_1^0 = m_1$ (and $m_2^0 = m_2$).

Thus invariance under translations, when combined with invariance under rotations and Lorentz boosts, requires the existence of a Poincaré invariant linear combination of the 4-velocities

$$u_s^\mu = (m_1 u_1^\mu + m_2 u_2^\mu)/N_s \tag{11.59}$$

that remains constant during the collision.

Exercise 11.4 *Write out the time component of this 4-vector equation and using the explicit form of the time components in terms of c, show that the normalization N_s is dictated to be that given in Eq.11.39 if the coefficients m_i/N_s are common to all components of u_s^μ.*

One comment is in order here. The coordinates $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ transform to locations at different times. Thus the 4-velocities after transformation are evaluated at different times. However, both before and after the collision, Newton's first law applies to each particle. Thus the 4-velocities of each particle are constant, except during the collision — where the spatial separation of the particles is zero and so no time differences appear after transformation to other inertial frames for this instant. Hence, the 4-velocities in Eq. 11.59 may be evaluated at equal times in every inertial frame.

11.4 Classical constraints on the coefficients m_i.

Since the coefficients in Eq. 11.59 are Poincaré invariant we may consider a frame in which $\mathbf{u}_s = 0$. In this frame we must have $m_1\mathbf{u}_1 = -m_2\mathbf{u}_2$. Therefore, either m_1 and m_2 are of the same sign while $\mathbf{u}_1$ and $\mathbf{u}_2$ are opposite, or m_1 and m_2 are opposite while $\mathbf{u}_1$ and $\mathbf{u}_2$ have the same sign. Let us assume the latter case, given that a collision does indeed take place. There are then two possibilities: either particle 1 passes through particle 2 or both particles reverse direction during the collision. Both possibilities are excluded by the classical particle axiom.

Hence, both m_1 and m_2 must have the same sign. This is an important conclusion and one that does not follow from Poincaré invariance alone.

11.5 Particle mass

The overall sign of m_1 and m_2 is still arbitrary. However, we can decide to adopt a convention in which it is always positive.

Exercise 11.5 *Show that in a two particle contact collision in which the particles retain their identity that m_2 and m_1 are related by*

$$m_2 = \frac{\gamma_1^i - \gamma_1^f}{\gamma_2^f - \gamma_2^i}\, m_1 \tag{11.60}$$

where the superscripts i and f refer to initial and final velocities respectively.

Choosing one of m_1 or m_2 sets the scale (and units) and thus fixes the other.

If we consider $(m_i u_i)^2 = m_i^2[(u_i^0)^2 - \mathbf{u}_i^2] = m_i^2 c^2$ we see that it is independent of $\mathbf{u}_i$. (One should compare this situation with that for the coefficients c_i which clearly depend on the particle speed.) This suggests that m_i is a property of the particle i rather than the collision. We are thus led to the following definition.

Definition 4 *Mass, m, is that Poincaré invariant property of a (classical) particle such that the sum of the combinations mu^μ of the mass and 4-velocities of each particle partaking in a collision is conserved. The actual value is fixed by reference to a chosen particle whose mass is arbitrarily chosen as a standard.*

For this definition to make sense we must show that if we have three particles, one of which is the standard, and we arrange for the other two to collide with each other and the standard, then the masses inferred from each of the three collisions are consistent. First we note that such a three-way collision must be coplanar and is possible in principle.

Exercise 11.6 *Show that $m_1 u_1^\mu + m_2 u_2^\mu + m_3 u_3^\mu$ is constant throughout the collision and infer that the consistency condition for mass is satisfied.*

11.6 More complicated interactions

The earlier discussion of collisions is easily extended to N particles. One finds that there are always $N - 1$ linearly independent internal coordinates and one system coordinate

$$\mathbf{x}_s = \sum_{i=1}^{N} m_i \mathbf{x}_i / N_s \tag{11.61}$$

whose corresponding 4-velocity

$$u_s^\mu = \sum_{i=1}^{N} m_i u_i^\mu / N_s \tag{11.62}$$

remains constant during the collision. (See problem 1.) The system coordinate is sometimes referred to as the system *center of mass*.

It is an experimental fact that in some collision processes, particles may disappear and be replaced by other particles. One might imagine this perhaps as one particle swallowing up another and taking on a new identity as a result. It is debatable whether a classical point particle can engage in such an interaction, though extended particles certainly can and do. For example, two balls of putty fired at each other stand a good chance of sticking together rather than bouncing off of each other. The classical particle axiom admits this possibility and our task now is to analyze it.

What takes place in such a collision process is not so much dependent on the particles participating as on the nature of the interaction. Knowledge of this interaction can only be gleaned from analysis of a collision. It must though be describable entirely in terms of the positions and velocities of the participating particles since this is the only information available. Consistency with translation invariance requires that any dependence on particle positions must be attributable entirely to changes in the internal difference coordinates. In the case of an interaction taking place at a point (as with a contact interaction) this means that the interaction cannot depend in any way on the collision point. As was the case when the number of particles was not altered, there must exist amongst the remaining linearly independent coordinates, one that describes the system as a whole in such a way that this system coordinate moves uniformly through the collision point. The difficulty now is that the description of this coordinate, in terms of mass weightings of the individual particle coordinates, must change as the number of particles changes. Nevertheless one notes that the particles before the collision cannot predict the future and the relevant system coordinate must therefore be the same as if there were to be no change in the particles during the collision. Hence the system coordinate must be the earlier x_s with the normalization of Eq. 11.39. Furthermore, it is possible to arrange for ordinary two-particle collisions to take place between a particle in the initial state and an additional (external) particle, which later undergoes another collision with a particle in the final state. Analysis shows that these collisions may be regarded as taking place with a composite system of total mass M_s given by Eq. 11.39 and thus the final system coordinate is also of the same form as before.

Review

Does Newton's First Law apply to individual particles at the instant of collision?

What is linear independence of vectors?

Why are the spatial coordinates of two separated particles at time t independent of each other?

What is configuration space and what is its dimension for N particles?

Does a particle's mass depend on its velocity?

Questions

1. Would the mass of a particle have any meaning if there were no interactions of any kind in the universe?

2. Suppose that two particles interact via other than a contact collision (e.g. an electrical or gravitational interaction). Is there a point of collision? Is there a vector analogous to $\mathbf{x}_c$?

3. Is there any reason why a two particle collision should be coplanar? Is a coplanar collision coplanar in every inertial frame?

4. Poincaré symmetry permits the existence of particles with negative mass. What properties might such particles exhibit (especially if they can coexist with particles of positive mass)?

5. The treatment of this chapter is somewhat deficient in concentrating on contact collisions. If one were to treat more general interactions, should one expect the same linear combinations of the coordinates to remain unaltered during the interactions? (In other words, are the particles' masses truely a property only of the particles or do they depend on the nature of the interactions?)

Problems

1. Consider a collision involving N particles.

   (a) Show that one can always choose $N(N-1)/2$ distinct pairs but that there are just $N-1$ linearly independent internal difference coordinates $\mathbf{x}_i - \mathbf{x}_j$. Conclude therefore that there must always exist one further linearly independent combination of the coordinates that changes at a constant rate throughout the collision.

   (b) Generalize the discussion leading to Eq. 11.59 to the case of N particles. (Note that it is not necessary to introduce the concept of mass for subsystems of particles in order to do this.)

2. Consider a glancing collision between two particles, as in fig. 11.3. The

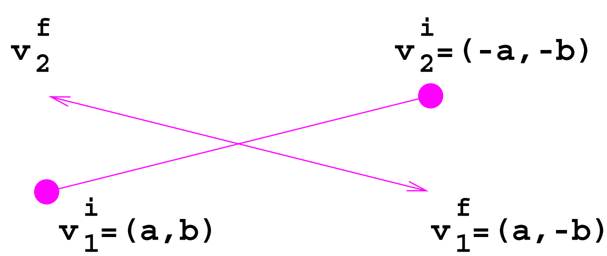

Figure 11.3: A glancing collision between two particles.

collision takes place in the xy plane and particle 1 has initial velocity $\mathbf{v}_1^i = (a, b)$ and final velocity $\mathbf{v}_1^f = (a, -b)$. Particle 2 is identical in every respect to particle 1 and has initial velocity $\mathbf{v}_2^i = (-a, -b)$, being equal in magnitude but opposite in direction to the initial velocity of particle 1.

(a) Find the initial velocity $\mathbf{v}_c$ and 4-velocity u^μ_s of the system making use of the fact that the particles are identical.

(b) What must the final velocity of particle 2 be if it is identical in nature to particle 1?

(c) Transform the various velocities to a frame moving in the x direction with speed $V = a$. Show that the transformed $\mathbf{v}_c$ does not remain constant during the collision (i.e. show that the same weighted average of the individual velocities does not remain constant). What components fail to remain constant?

(d) Calculate the 4-velocities of the two particles in this new frame and show that each component of u^μ_s remains constant.

3. Show that if one takes the vector $\mathbf{x}_c$ defined in one inertial frame and boosts it to another inertial frame then the resultant vector is not the same as the vector $\mathbf{x}_c$ defined directly in the new frame.

Chapter 12

Momentum and Energy

The conserved combination of mass and 4-velocity is so useful that we give it its own name: 4-momentum. Thus the 4-momentum is defined as

$$p^\mu = mu^\mu = m\frac{dx^\mu}{d\tau}. \tag{12.1}$$

Using the connection between 4-velocity and ordinary velocity,

$$u^\mu = \gamma(c, \mathbf{v}), \tag{12.2}$$

stemming from the relationship between ordinary time and proper time, we obtain

$$p^\mu = (p^0, \mathbf{p}) = m\gamma(c, \mathbf{v}). \tag{12.3}$$

The quantity

$$\mathbf{p} = \gamma m\mathbf{v} \tag{12.4}$$

is called simply the *momentum*.

When more than one particle is present, the total momentum of the system is defined as the sum of the momenta of its constituent particles.

12.1 Transformations for momentum

The transformation equations for the 4-momentum follow immediately from those of the 4-velocity and like the 4-velocity are identical in form to those of the spacetime coordinates x^μ.

If we change to a frame moving with speed V along the positive x axis then the 4-momentum in this frame is just

$$p'^{\mu} = mu'^{\mu} = m\frac{dx'^{\mu}}{d\tau}. \tag{12.5}$$

Using the Poincaré invariance of m and the transformations for the components of the 4-velocity, shows that

$$\begin{aligned} p'^{0} &= \gamma_V(p^0 - \beta_V p^x) & (12.6)\\ p'^{x} &= \gamma_V(p^x - \beta_V p^0) & (12.7)\\ p'^{y} &= p^y & (12.8)\\ p'^{z} &= p^z. & (12.9) \end{aligned}$$

Exercise 12.1 *Show that the total momentum $P^\mu = \sum_i p_i^\mu$ of a system of particles obeys the same transformation equations. Hence confirm that if each component of P^μ is conserved (during collisions) in one frame then it is conserved in every frame.*

12.2 The conservation law

Although the conservation of 4-momentum is little more than a product of our construction of mass and momentum it reflects the underlying translation symmetry of physics. With independent knowledge of a particle's mass, the conservation of 4-momentum also becomes a powerful analysis tool. Results as general as the conservation of 4-momentum assume an important role in physics and prompt us to give them a special name.

Definition 5 *A law of physics is a predictive description of a class of physical phenomena that is independent of the choice of inertial reference frame.*

Since the conservation of 4-momentum applies in any inertial frame, and places constraints on the outcome of a collision experiment, it is classed as a law of physics. It is in fact one of the most important laws of physics. So great is our faith in the conservation law that when experiments appeared to show that 4-momentum was not conserved in the so-called β-decay of neutrons in atomic nuclei, it was proposed that there must exist a hithertofore unknown particle, termed the neutrino, that carried away the missing 4-momentum. Subsequent investigations have confirmed the existence of this extremely difficult-to-detect particle.

12.3 Energy

The interpretation of the time component, p^0, of the 4-momentum is aided by consideration of the non-relativistic limit. In the limit that $v \ll c$, one has

$$\gamma = (1 - \frac{v^2}{c^2})^{-1/2} \approx (1 + \frac{v^2}{2c^2} + \cdots). \tag{12.10}$$

Retaining terms of order v^2/c^2 only, we find

$$\mathbf{p} = \gamma m \mathbf{v} \approx m\mathbf{v}, \tag{12.11}$$

which is the Newtonian expression for the momentum (thus justifying the name we have given it), while

$$p^0 = \gamma mc \approx mc + \frac{1}{2}mv^2\frac{1}{c} + \cdots . \tag{12.12}$$

The quantity $mv^2/2$ is the Newtonian kinetic energy. Thus

$$p^0 c = mc^2 + \frac{1}{2}mv^2 + \cdots \tag{12.13}$$

has the same units as Newtonian energy. We therefore define the *energy* of a particle, in a relativistic context, as

$$E = p^0 c = \gamma mc^2. \tag{12.14}$$

One notes that if a particle is at rest, then $\gamma = 1$, and the energy in this frame is

$$E_0 = mc^2. \tag{12.15}$$

E_0 is called the *rest energy*. This equation is Einstein's famous mass-energy equivalence. It shows that mass is a form of energy. It is worth noting that this relationship between mass and energy was not recognized in Newtonian physics.

The energy and momentum can be used to provide some useful formulae for relativistic factors. Firstly, we obtain immediately from inversion of Eq. 12.14 that

$$\gamma = \frac{E}{mc^2} = \frac{E}{E_0}. \tag{12.16}$$

We know that if γ is close to 1 then non-relativistic approximations may be used. Thus the validity of non-relativistic theory may be equivalently checked by comparing the total energy and the rest energy.

Taking the ratio of Eqs. 12.4 and 12.14 yields

$$\frac{\mathbf{p}c}{E} = \beta \tag{12.17}$$

from which one deduces that the particle velocity is given by

$$\mathbf{v} = \frac{\mathbf{p}c^2}{E}. \tag{12.18}$$

Notice that in the ultrarelativistic limit, $v \to c$, we have that the magnitude of the momentum is given by

$$p = \frac{E}{c}. \tag{12.19}$$

12.4 A Poincaré invariant

Since

$$(u^0)^2 - \mathbf{u}^2 = \gamma^2(c^2 - v^2) = c^2 \tag{12.20}$$

we find that

$$p^2 \equiv p \cdot p \equiv (p^0)^2 - \mathbf{p}^2 = m^2c^2. \tag{12.21}$$

Exercise 12.2 *Show by explicit calculation that the left hand side of this expression is Poincaré invariant, i.e. that it is invariant under Lorentz boosts, rotations and spacetime translations.*

The Poincaré invariance of this expression is to be compared to the spacetime interval for the coordinates, which is only Lorentz invariant and is not invariant under translations.

Another way of expressing the above result, which is more widely used, is to substitute $p^0 = E/c$. This leads to

$$E^2 - p^2c^2 = m^2c^4. \tag{12.22}$$

This equation is true for any particle that is free of external influence.

Some important conclusions can be drawn from this invariant. Firstly, we note that, for real masses, one must always have $E^2 > p^2c^2$. Suppose though that the mass of the particle were zero. Then we must have

$$p = E/c. \tag{12.23}$$

In this case, it follows from Eq. 12.18 that the particle velocity must be c. Conversely, if the particle velocity is c then application of Eq. 12.18 in Eq. 12.22 shows that it must be massless. One might well quibble that our earlier derivations are not applicable to massless particles. Nevertheless, the above statements hold in the limit of infinitesimal mass. A more careful analysis of things such as light, which do travel at c, shows that it is meaningful to consider them as conveying momentum and energy, which are conserved in interactions with massive particles, and which are related by Eq. 12.23.

Exercise 12.3 *Use Eq. 12.22 to show that in the particle's rest frame*

$$E = E_0 \equiv mc^2.$$

It is amusing to consider the fact that energy, momentum and mass all appear in Eq. 12.22 as squared quantities. If one tries to determine the energy of a particle from its known momentum and mass, one finds

$$E = \pm\sqrt{p^2c^2 + m^2c^4}. \tag{12.24}$$

Mathematically, both positive and negative square roots are permitted. This is a hint of the existence in relativistic theories of both particles and antiparticles. Antiparticles correspond to the negative energy solution, though there are ways of reinterpreting them as positive energy particles.

12.5 Kinetic energy

The *kinetic energy* is defined as

$$K \equiv E - E_0 = (\gamma - 1)mc^2. \tag{12.25}$$

Exercise 12.4 *Show that in the low velocity limit, $v \ll c$,*

$$K \approx \frac{1}{2}mv^2. \tag{12.26}$$

A useful relation between kinetic energy and momentum can be obtained by manipulating Eq. 12.22. We have

$$E^2 - p^2c^2 = E_0^2, \tag{12.27}$$

and rearranging,

$$E^2 - E_0^2 = p^2c^2. \tag{12.28}$$

Now observe that the left-hand side is

$$E^2 - E_0^2 = (E - E_0)(E + E_0) = K(E + E_0). \tag{12.29}$$

Hence,

$$\begin{aligned} K &= \frac{p^2c^2}{E + E_0} \\ &= \frac{p^2c^2}{\gamma mc^2 + mc^2} \\ &= \frac{p^2}{(\gamma + 1)m}. \end{aligned} \tag{12.30}$$

In the non-relativistic limit this reduces to $p^2/2m$.

12.6 The mass of composite systems

Application of the conservation laws of energy and momentum in many situations requires an understanding of what is meant by the mass of a composite system. Since the total energy and momentum of a composite system is defined as the sum of the energies and momenta of its constituents (see exercise 12.1) we should define the mass of the system via Eq. 12.22.

In a frame in which the total momentum of the system is zero, we obtain

$$M = E_0/c^2 \tag{12.31}$$

where

$$E_0 = \sum_i E_i^{\text{system rest frame}}. \tag{12.32}$$

Since the constituents will usually be moving in the rest frame of the system, their individual energies will contain both kinetic and mass terms. Thus the mass of the system will, in the absence of all but contact interactions, be greater than the sum of the masses of its constituents.

Exercise 12.5 *Show that in a frame moving with respect to the system rest frame with a speed corresponding to the time dilation factor γ_s that the system mass is just the normalization adopted in Eq. 11.39.*

12.6.1 Binding energy

Sometimes there are interactions present which may result in the constituents being bound together, e.g. the nuclear interaction may bind protons and neutrons together in the atomic nucleus, the electromagnetic interaction may bind the electrons and nuclei of atoms together and the gravitational interaction may bind the planets and the Sun. In general, binding requires that an interaction be attractive (at least over the relevant distance range).

Often it is possible, especially when the constituents are bound together, to determine the momentum and energy, and hence the mass, of a composite system without having to sum up the energy and momenta of its constituents. For example, atomic nuclei, atoms and molecules will often participate in interactions with other systems as a coherent whole. Thus the system energy and momentum can be determined.

It is found that for bound systems, the mass of the system is always less than the sum of the masses of its constituents. this leads us to define the *mass deficit* as

$$\Delta m = \sum_i m_i - M_{system}. \tag{12.33}$$

The corresponding energy difference,

$$\Delta E = \Delta mc^2 \tag{12.34}$$

is known as the *binding energy.*

> **Example** The deuteron is the nucleus of the isotope of hydrogen known as deuterium. It consists of one proton and one neutron. In dealing with subatomic systems it is common to use atomic mass units, denoted by u. These have the value
>
> $$1\,u = 931.50\,\mathrm{MeV/c^2}. \tag{12.35}$$
>
> An MeV is another common unit in dealing with subatomic systems. It is 10^6 "electron volts" (eV) where an electron volt is

the energy that an electron would acquire upon being accelerated through an electric potential of 1 volt. In these units one finds that the masses of the proton and neutron are

$$m_p = 1.007276\,u \tag{12.36}$$
$$m_n = 1.008665\,u \tag{12.37}$$

whose sum is

$$m_p + m_n = 2.015941\,u. \tag{12.38}$$

But the mass of the deuteron is

$$m_d = 2.013553\,u. \tag{12.39}$$

Therefore the mass deficit is

$$\Delta m = (m_p + m_n) - m_d = 0.002388\,u \tag{12.40}$$

and so the binding energy is

$$\begin{aligned} \Delta E &= \Delta m c^2 \\ &= 0.002388\,u \times 931.50 \frac{\text{MeV}/u}{c^2} \times c^2 \\ &= 2.22\,\text{MeV}. \end{aligned} \tag{12.41}$$

This binding energy is actually quite small. The average binding energy per nucleon (i.e. proton or neutron) in a typical atomic nucleus is about 8 MeV. So, comparatively speaking, the deuteron is quite weakly bound.

When a proton and neutron fuse to form deuterium the binding energy is released. This is the basis for the hydrogen bomb. When one has an enormous number of nuclei, as in macroscopic amounts of matter, the total energy released is huge. Similar processes fuel the Sun. For example, if two deuterons fuse to form a ^{4}He nucleus then 23.9 MeV of energy is released. The basic energy source for the Sun is the nuclear conversion of hydrogen to helium and this is one of the reactions involved. The enormous supply of hydrogen in the Sun and the large amount of energy available from this source, enables it to continue shining for billions of years. Until the discovery of nuclear reactions, the source of this amount of energy had been a mystery.

In contrast, the binding energy of the electron in the ground state of hydrogen is only 13.6 eV. This is typical of the electrons in atoms and is why the energies involved in chemical reactions (which involve atomic electrons) are so much less than those in nuclear reactions.

12.7 Consequences and applications of energy-momentum conservation

Most of the examples discussed in this section deal with subatomic particles. Strictly speaking, these are not classical particles but are complex, and rather puzzling, objects called quanta. However, they do exhibit many classical particle properties and indeed the success of the treatments presented here is compelling evidence for that. Of particular note is that we shall treat light as composed of massless particles called photons.

12.7.1 Relativistic billiards

Consider a particle (e.g. a proton, electron or billiard ball) which collides elastically with an identical stationary particle, of the same mass, m. (An elastic collision means one in which kinetic energy is conserved, so no energy is lost to internal excitations or omission of radiation.) See fig. 12.1. By

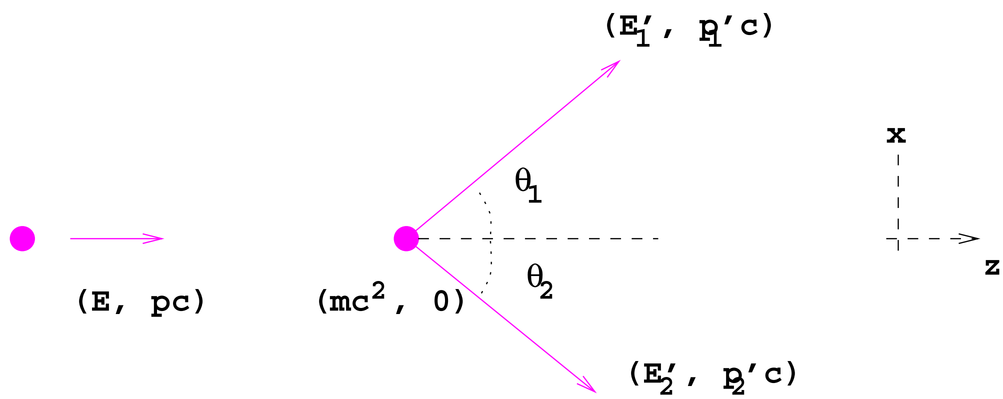

Figure 12.1: A particle is scattered by an identical particle initially at rest. The scattering takes place in the zx-plane.

conservation of energy:

$$E + mc^2 = E_1' + E_2'.$$

By conservation of momentum:

$$\begin{aligned} 0 &= p_{1x}' + p_{2x}' \\ &= p_1' \sin\theta_1 - p_2' \sin\theta_2 \end{aligned}$$

and

$$\begin{aligned} p_z &= p_1' \cos\theta_1 - p_2' \cos\theta_2 \\ &\equiv p. \end{aligned}$$

Consider the case

$$\theta_2 = \theta_1 \equiv \theta/2.$$

Then

$$p_1' = p_2' \equiv p'$$

and therefore

$$p = 2p' \cos\frac{\theta}{2}.$$

It follows that

$$E_1' = E_2' \equiv E' = \sqrt{p'^2c^2 + m^2c^4}$$

and from the conservation law

$$E' = \frac{E + mc^2}{2}.$$

In order to proceed, we note the trigonometric identity

$$\cos\theta = \cos^2\frac{\theta}{2} - \sin^2\frac{\theta}{2} = 2\cos^2\frac{\theta}{2} - 1.$$

Thus

$$\begin{aligned} \cos\theta &= 2\frac{p^2}{4p'^2} - 1 \\ &= \frac{p^2c^2}{2p'^2c^2} - 1. \end{aligned}$$

Noting Eqs. 12.28 and 12.29, we have

$$p^2c^2 = K(E + E_0) = K(K + 2E_0).$$

Similarly,

$$\begin{aligned}
p'^2c^2 &= E'^2 - E_0^2 \\
&= \left(\frac{E+E_0}{2}\right)^2 - E_0^2 \\
&= \frac{(K+2E_0)^2}{4} - E_0^2 \\
&= \frac{K}{4}(K+4E_0).
\end{aligned}$$

Hence

$$\begin{aligned}
\cos\theta &= \frac{K(K+2E_0)}{2\frac{K}{4}(K+4E_0)} - 1 \\
&= \frac{K}{K+4E_0}.
\end{aligned}$$

In the low velocity limit, the kinetic energy K is much smaller than the rest energy E_0. It can therefore be neglected in comparison and we can consider the limit $K \to 0$. Then, $\cos\theta \to 0$. This implies that the angle θ between the scattered particles is $90°$, a result well-known to billiard (and snooker and pool) players.

On the other hand, as the kinetic energy becomes very large and one approaches the ultrarelativistic regime, $K \to \infty$, we see that $\cos\theta \to 1$. This implies that the angle θ tends to $0°$, i.e. the scattering angle between the particles is squashed at high energies. This may be compared with the headlight effect. Observations of particle reactions in bubble chambers first revealed this phenomenon as early as 1932 and it was considered then as an important confirmation of relativistic mechanics.

12.7.2 Compton scattering

Compton scattering is the scattering of a photon, such as an x-ray or gamma ray, by a massive particle such as an electron, usually taken to be at rest. See fig. 12.2. We denote the energy of the incident photon by Q and that of the scattered photon by Q'. The momenta of the photons have magnitudes given by Eq. 12.23. We denote the directions of these momenta by the unit vectors $\mathbf{n}$ and $\mathbf{n}'$ respectively. The recoil electron has energy E and momentum $\mathbf{p}$.

By conservation of energy:

$$Q + m_e c^2 = E + Q'$$

and by conservation of momentum

$$\frac{Q}{c}\mathbf{n} = \frac{Q'}{c}\mathbf{n}' + \mathbf{p}.$$

Simple manipulations lead to the pair of equations:

$$\begin{aligned}(Q - Q') + m_e c^2 &= E \\ Q\mathbf{n} - Q'\mathbf{n}' &= \mathbf{p}c.\end{aligned}$$

Squaring them gives

$$\begin{aligned}E^2 &= (Q - Q')^2 + 2(Q - Q')m_e c^2 + (m_e c^2)^2 \\ p^2 c^2 &= Q^2 - 2QQ' \cos\theta + Q'^2\end{aligned}$$

and subtracting these must yield the Poincaré invariant $m_e^2 c^4$:

$$\begin{aligned}2QQ'(1 - \cos\theta) - 2(Q - Q')m_e c^2 + m_e^2 c^4 &= m_e^2 c^4 \\ \Longrightarrow (Q - Q')m_e c^2 &= QQ'(1 - \cos\theta).\end{aligned}$$

Dividing through by $QQ'm_e c^2$ leads to the Compton scattering formula

$$\frac{1}{Q'} - \frac{1}{Q} = \frac{1}{m_e c^2}(1 - \cos\theta).$$

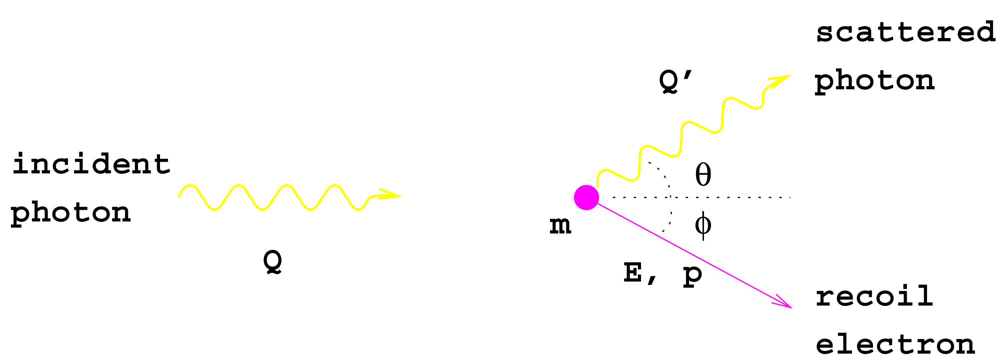

Figure 12.2: Compton scattering of a photon by a stationary electron

Since the right hand side of this equation is always positive, the scattered photon has a lower energy than the incident photon. (This manifests itself as a lower frequency.) When Compton discovered this effect in 1922, and found that it was well described by the above formula, it became a celebrated confirmation that light exhibits particle properties.

12.7.3 Annihilation

The neutral pion, π^0, is a particle with a mass of approximately 140 MeV$/c^2$. It is unstable and its dominant decay mode is into two gamma rays. Consider a π^0 at rest in the laboratory frame. See fig. 12.3. By conservation of energy

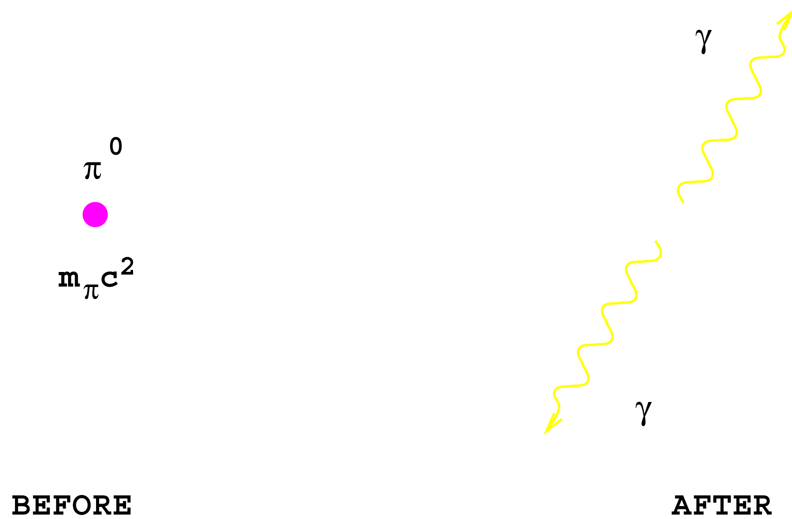

Figure 12.3: Decay of a neutral pion at rest.

$$m_\pi c^2 = E_{\gamma_1} + E_{\gamma_2}$$

and by conservation of momentum

$$0 = \mathbf{p}_{\gamma_1} + \mathbf{p}_{\gamma_2}.$$

Thus

$$\mathbf{p}_{\gamma_1} = -\mathbf{p}_{\gamma_2} \equiv \mathbf{p}_\gamma$$

and, since $E = pc$ for a (massless) photon,

$$E_{\gamma_1} = E_{\gamma_2} \equiv E_\gamma.$$

Thus

$$\begin{aligned} E_\gamma &= \frac{1}{2} m_\pi c^2 \\ &= 70\,\mathrm{MeV}. \end{aligned}$$

In the quark model of elementary particles, the π^0 is a quark-antiquark pair in a bound state. The decay thus represents the annihilation of the quark-antiquark pair.

We observe that in an annihilation process such as this one, all mass is converted to radiation energy. It is important to understand what is meant by this statement though. It simply refers to the nature of the constituent energy of the system. The two photon final state can still be ascribed a mass of 140 MeV/c^2, since the pair has zero net momentum and the mass of the system is given by Eq. 12.31.

12.7.4 Photon Integrity

One interesting consequence of the conservation laws of energy and momentum is that the combination of them prevents a photon from breaking up into two photons (unless they both move in the same direction as the original). See fig. 12.4. Consider the conservation of energy and momentum:

$$\begin{aligned} E &= E_1 + E_2 \\ \mathbf{p} &= \mathbf{p}_1 + \mathbf{p}_2. \end{aligned}$$

The second of these means that the momentum vectors must form a triangle to accomplish the vector addition. From the triangle inequality the magnitudes of the momenta satisfy

$$p \leq p_1 + p_2$$

i.e., the length of one side of the triangle is less than the sum of the lengths of the other two sides. The equality can only occur if two sides collapse onto the third, the triangle then having no area. However, for a photon, $p = E/c$ and so energy conservation implies that

$$p = p_1 + p_2$$

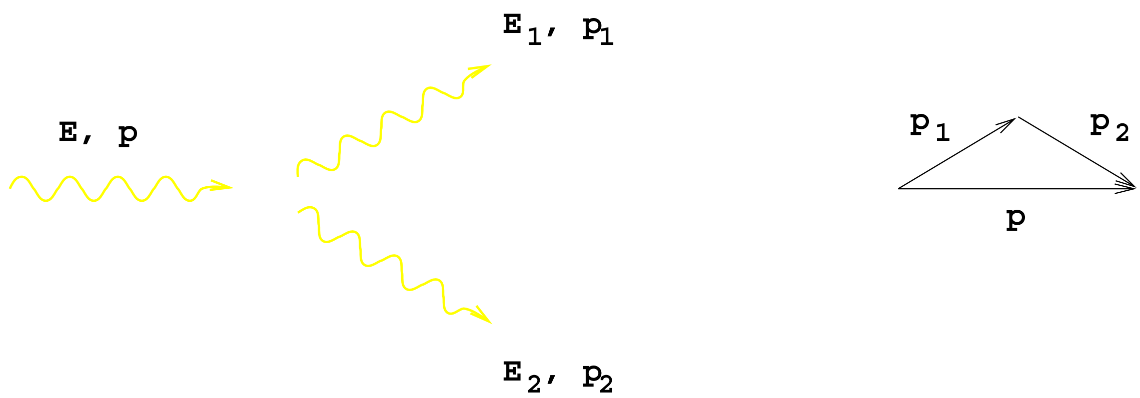

Figure 12.4: A photon cannot decay. The photon momenta must form a triangle and this is inconsistent with energy conservation for massless particles.

leading to a contradiction, unless both $\mathbf{p}_1$ and $\mathbf{p}_2$ are parallel to $\mathbf{p}$. Actually, even this can be excluded in the standard quantum theory of light.

12.7.5 Absorption and emission of radiation by atoms and nuclei

Atoms and nuclei emit radiation with a discrete spectra. This can be understood as due to transitions between discrete energy levels within the atom or nucleus. Similarly, atoms and nuclei can absorb radiation and make a transition to an excited state. Of course, the excited states have different masses due to the binding energies involved.

Consider the emission of radiation from an atom (in an excited state) at rest. To conserve momentum, the final state atom must recoil away from the emitted photon. See fig. 12.5. Thus the energy of the emitted photon must be less than the excitation energy of the atom (i.e. the difference in energy between the initial and final atomic states).

On the other hand, if an atom at rest absorbs radiation, the excited atom must recoil to conserve momentum. Therefore the energy of the absorbed photon must be greater than the excitation energy of the atom.

It follows that if atomic energy levels were perfectly sharp, an atom could not absorb its own radiation. However, it is readily observed that if the radiation from a gas discharge tube is passed through another tube of the

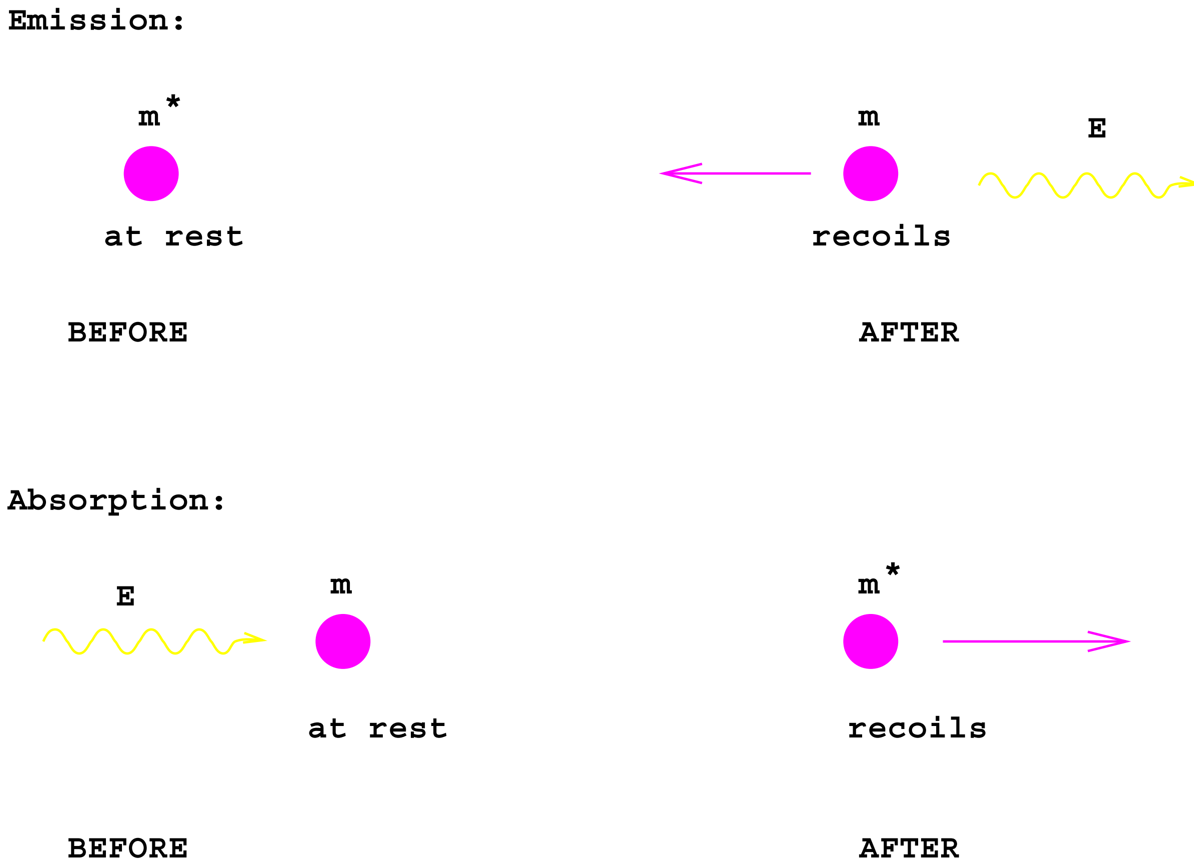

Figure 12.5: Emission and absorption of radiation by atoms.

same gas, then absorption of radiation occurs. This may be attributed to the finite widths of the spectral lines. Most of this width is due to doppler broadening associated with the thermal motion of the atoms.

However, if the photon energy is large enough, as for example in a gamma ray emitted from an atomic nucleus, then the recoil becomes more pronounced and the effect is observable. It was first seen in 1951 using the Mossbauer effect.

Similar considerations play a role in the laser cooling of atoms. If (laser) light of the right frequency is absorbed by an atom headed towards the oncoming light then conservation of energy and momentum will slow it down. Since an atom moving away from the laser light will absorb at a slightly different frequency it will not absorb the same light and so will not be speeded up in the opposite direction. Though the atom will recoil when it reemits the light, that reemission will be in all directions. Hence, on average, atoms moving towards the laser will be slowed down, in that direction. By using 6 lasers, 3 pairs at right angles and the lasers in each pair opposite to each other, the motion of atoms in all directions can be slowed, i.e. the atom is cooled. This mechanism has been used to cool atoms to extremely low temperatures.

Review

What is the relativistic expression for the momentum of a particle of mass m?

What is the relativistic expression for the (total) energy of a particle of mass m?

Write down a Poincaré invariant expression relating energy and momentum.

Explain briefly what Compton scattering is, paying attention to the role of energy conservation in the process.

A particle-antiparticle pair has a minimum energy of $2mc^2$. If a lone photon has at least this much energy can it create this pair? If not, why not.

Questions

1. What properties might you expect for a particle with imaginary mass (i.e. a particle whose mass squared is negative)?

2. A gas consists of molecules in random motion. As the temperature is increased the average velocity of the gas molecules increases. Does the mass of a container of gas depend on the temperature? If so, is it a significant effect? (Typical molecular velocities at room temperature are of the order of 1000 m/s.)

3. Is the mass of a compressed spring the same as that of the same spring with normal extension?

Problems

1. Show that a neutral pion cannot decay into a single photon.

2. Show that a lone photon cannot decay into an electron-positron pair in the absence of other matter or radiation. (The electron, e^-, and the positron, e^+, are a particle-antiparticle pair with the same (non-zero) mass as each other.)

3. A particle of mass 2 MeV/c^2 and kinetic energy 3 MeV collides with a stationary particle of mass 4 MeV/c^2. After the collision, the two particles stick together. Find

   (a) the initial momentum of the system,

   (b) the mass of the fused pair in the final state.

4. A particle of mass 770 MeV/c^2 decays at rest in the laboratory frame into two particles of mass 140 MeV/c^2 each. Determine the kinetic energy and momenta of each of the decay products. Determine also the velocities of the decay products.

Chapter 13

Force

In contact collisions, and other situations, individual particles do change their momentum. We can speak of this change by saying that a *force* has acted on the particle.

13.1 Newton's second law

We define a force, $\mathbf{F}$, as the time rate of change of a particle's momentum, $\mathbf{p}$, i.e.

$$\mathbf{F} \equiv \frac{d\mathbf{p}}{dt}. \tag{13.1}$$

This equation is known as Newton's second law. In the present approach it should be appreciated that this terminology arises for historical reasons. Newton's approach to mechanics was logically different from ours, where the equation is a mere definition rather than a true law.

Consider now, the evaluation of Eq. 13.1 for $\mathbf{p} = \gamma m\mathbf{v}$. Since the particle's mass m is constant we have

$$\begin{aligned}
\mathbf{F} &= m\frac{d(\gamma\mathbf{v})}{dt} \\
&= m\gamma\frac{d\mathbf{v}}{dt} + m\mathbf{v}\frac{d\gamma}{dt} \\
&= \gamma m\mathbf{a} + m\mathbf{v}\frac{d\gamma}{dt}
\end{aligned} \tag{13.2}$$

where

$$\mathbf{a} \equiv \frac{d\mathbf{v}}{dt} \tag{13.3}$$

is defined as the particle's *acceleration*. In the non-relativistic limit one has $\gamma = 1$ and the last term vanishes, leaving the well-known Newtonian result: $\mathbf{F} = m\mathbf{a}$. However, this is not generally true!

To obtain the full relativistic expression relating force and acceleration we must consider how γ varies with time. That it does so follows immediately from the dependence of γ on the particle velocity. In detail, we have

$$\begin{aligned}
\frac{d\gamma}{dt} &= \frac{d}{dt}\left(1 - \frac{v^2}{c^2}\right)^{-1/2} \\
&= -\frac{1}{2}\left(1 - \frac{v^2}{c^2}\right)^{-3/2}\left(-\frac{1}{c^2}\right)\frac{dv^2}{dt} \qquad (13.4) \\
&= \frac{\gamma^3}{2c^2}\frac{dv^2}{dt}. \qquad (13.5)
\end{aligned}$$

Now observe that

$$\frac{dv^2}{dt} = \frac{d\mathbf{v}\cdot\mathbf{v}}{dt} = 2\mathbf{v}\cdot\frac{d\mathbf{v}}{dt} = 2\mathbf{v}\cdot\mathbf{a} \tag{13.6}$$

where "·" means a dot, or scalar, product of vectors. Thus,

$$\frac{d\gamma}{dt} = \gamma^3\mathbf{v}\cdot\mathbf{a}/c^2. \tag{13.7}$$

Therefore

$$\mathbf{F} = \gamma m\mathbf{a} + \frac{\gamma^3\mathbf{v}\cdot\mathbf{a}}{c^2}m\mathbf{v}. \tag{13.8}$$

It is seen from this equation that in general $\mathbf{F}$ and $\mathbf{a}$ are not parallel.

Exercise 13.1 *Show that (except in the special case* $\mathbf{v} = 0$*) the only circumstances in which* $\mathbf{F}$ *is in the same direction as* $\mathbf{a}$ *are*

1. *if* $\mathbf{v} \parallel \mathbf{a}$*, in which case* $\mathbf{F} = \gamma^3 m\mathbf{a}$
2. *if* $\mathbf{v} \perp \mathbf{a}$*, in which case* $\mathbf{F} = \gamma m\mathbf{a}$*.*

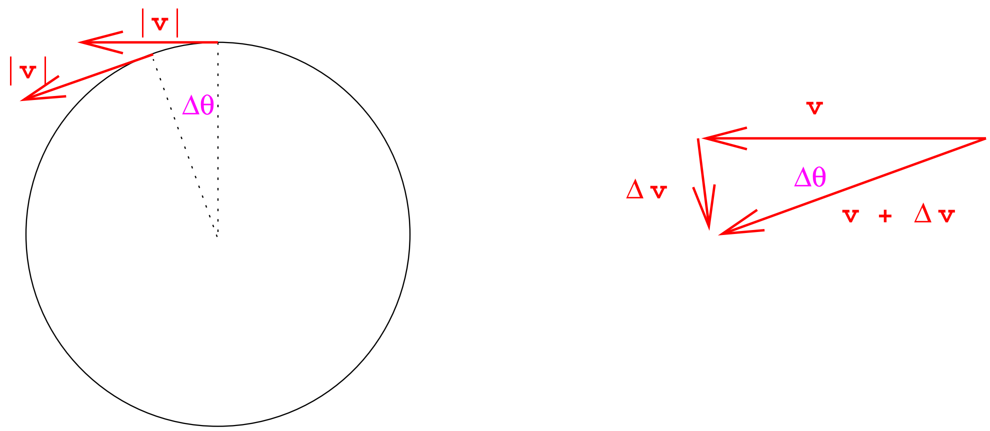

Figure 13.1: In the limit that $\Delta\theta$ becomes infinitesimally small, the change in velocity, $\Delta\mathbf{v}$, for uniform circular motion, is perpendicular to the instantaneous velocity, $\mathbf{v}$.

13.1.1 Centripetal forces

A simple instance of where the velocity and acceleration are perpendicular occurs for uniform motion in a circle. For a particle moving in a circle with uniform speed, the instantaneous velocity vector is always tangential to the circle. In order to keep the velocity vectors tangential, the instantaneous acceleration must be directed towards the center of the circle and thus it is called the *centripetal* acceleration. See fig. 13.1. However, it is a result of Euclidean geometry that the tangent and radius of a circle are at right angles to each other and hence the velocity and centripetal acceleration are perpendicular to each other.

In the limit that $\Delta\theta$ becomes very small, $\Delta\mathbf{v}$ becomes the arc length of an infinitesimal segment of a circle of radius v. Thus

$$dv = v d\theta \tag{13.9}$$

and the instantaneous acceleration is therefore

$$a = \frac{dv}{dt} = v\frac{d\theta}{dt} = v\omega \tag{13.10}$$

where, by definition,

$$\omega \equiv \frac{d\theta}{dt} \tag{13.11}$$

is the *angular velocity.*

If the circle is of radius r, then the length of an infinitesimal arc on the circumference is $ds = rd\theta$ and thus we have that the speed is

$$v = \frac{ds}{dt} = r\frac{d\theta}{dt} = r\omega. \tag{13.12}$$

Thus

$$\omega = v/r \tag{13.13}$$

and so the centripetal acceleration is (from Eq. 13.10)

$$a_{\text{centripetal}} = \frac{v^2}{r}. \tag{13.14}$$

It follows then from Eq. 13.8, in the case that $\mathbf{v}$ and $\mathbf{a}$ are perpendicular, that the force required to maintain the circular motion is in the direction of the centripetal acceleration and has magnitude

$$F_{\text{centripetal}} = \gamma m a_{\text{centripetal}} = \gamma m \frac{v^2}{r}. \tag{13.15}$$

We call it the centripetal force.

13.2 Newton's third law

Consider two particles colliding with each other and free of any influence external to the two particle system. Then, conservation of momentum applies and we have

$$\mathbf{p}_1 + \mathbf{p}_2 = \text{constant}. \tag{13.16}$$

Differentiating with respect to time, we have

$$\frac{d\mathbf{p}_1}{dt} + \frac{d\mathbf{p}_2}{dt} = 0 \tag{13.17}$$

But $d\mathbf{p}_1/dt$ is just the force, $\mathbf{F}_{12}$, on particle 1 due to particle 2 and $d\mathbf{p}_2/dt$ is the force, $\mathbf{F}_{21}$, on particle 2 due to 1. Thus

$$\mathbf{F}_{12} = -\mathbf{F}_{21} \tag{13.18}$$

i.e. the forces that the particles exert on each other are equal in magnitude and opposite in direction.

This result is known as Newton's third law.

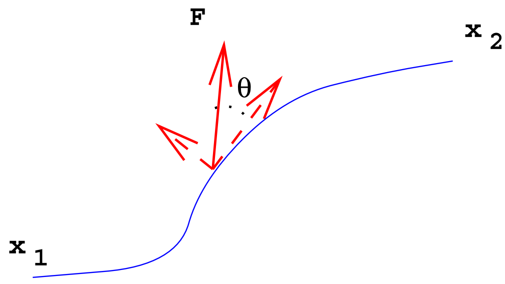

Figure 13.2: The work done by a force **F** on a particle depends on the component of force along the direction of the path travelled by the particle.

13.3 Work-Energy theorem

The *work* done by a force on a particle is defined as

$$W = \int_{x_1}^{x_2} \mathbf{F} \cdot d\mathbf{x}. \tag{13.19}$$

This involves a vector line integral. One must evaluate the scalar product

$$\mathbf{F} \cdot d\mathbf{x} = F \cos\theta \, dx = F_1 dx^1 + F_2 dx^2 + F_3 dx^3 \tag{13.20}$$

at all points along the path travelled by the particle between x_1 and x_2. See fig. 13.2. In general, the work done will depend on the path taken.

Consider the integrand. We have, using Eq. 13.2, that

$$\mathbf{F} \cdot d\mathbf{x} = \gamma m \mathbf{a} \cdot d\mathbf{x} + m\mathbf{v} \cdot d\mathbf{x} \frac{d\gamma}{dt}. \tag{13.21}$$

Now observe that

$$\mathbf{a} \cdot d\mathbf{x} = \frac{d\mathbf{v}}{dt} \cdot d\mathbf{x} = d\mathbf{v} \cdot \frac{d\mathbf{x}}{dt} = d\mathbf{v} \cdot \mathbf{v} = \frac{1}{2} d(\mathbf{v} \cdot \mathbf{v}) = \frac{1}{2} dv^2 \tag{13.22}$$

and similarly

$$\mathbf{v} \cdot d\mathbf{x} \frac{d\gamma}{dt} = \mathbf{v} \cdot \frac{d\mathbf{x}}{dt} d\gamma = v^2 d\gamma. \tag{13.23}$$

Inverting Eq. 13.5 we see that the first of these two equations can be written as

$$\mathbf{a} \cdot d\mathbf{x} = \frac{c^2}{\gamma^3} d\gamma. \tag{13.24}$$

Combining these results we thus obtain

$$\begin{aligned}\mathbf{F}\cdot d\mathbf{x} = \gamma m \frac{c^2}{\gamma^3} d\gamma + mv^2 d\gamma &= mc^2(\gamma^{-2} + \beta^2) d\gamma \\ &= mc^2 d\gamma. \end{aligned} \tag{13.25}$$

Therefore, the work done is

$$\begin{aligned} W = mc^2 \int_{\gamma_1}^{\gamma_2} d\gamma &= mc^2(\gamma_2 - \gamma_1) \\ &= \gamma_2 mc^2 - \gamma_1 mc^2. \end{aligned} \tag{13.26}$$

If the initial velocity is zero and the final velocity v then we have

$$W = \gamma mc^2 - mc^2 = E - E_0 \equiv K. \tag{13.27}$$

Thus we see that in general, the work done by the force is

$$W = \Delta K \tag{13.28}$$

where ΔK is the kinetic energy acquired by the particle. This result is known as the work-energy theorem.

13.4 Power

Power is defined as the time rate of change of a particle's energy, i.e.

$$P \equiv \frac{dE}{dt}. \tag{13.29}$$

Common experience tells us that the faster one expends energy doing work then the more effort is required. Thus power is a useful concept and a variety of machines (from motors to microwave ovens) are commonly rated by the power they can generate or consume.

If we manipulate the Poincaré invariant of Eq. 12.22 we obtain

$$E^2 = p^2c^2 + m^2c^4 = \mathbf{p}\cdot\mathbf{p}c^2 + E_0^2. \tag{13.30}$$

Differentiating both sides with respect to time we obtain

$$\begin{aligned} \frac{dE^2}{dt} = 2E\frac{dE}{dt} &= c^2 \frac{d(\mathbf{p}\cdot\mathbf{p})}{dt} + 0 \\ &= c^2 2\mathbf{p}\cdot\frac{d\mathbf{p}}{dt} \\ &= 2c^2\mathbf{p}\cdot\mathbf{F}. \end{aligned} \tag{13.31}$$

Thus

$$\frac{dE}{dt} = \mathbf{F} \cdot \mathbf{p}c^2/E. \tag{13.32}$$

Recalling Eq. 12.18 we have that the power is given by

$$P = \mathbf{F} \cdot \mathbf{v} \tag{13.33}$$

where $\mathbf{v}$ is the velocity of the particle on which the force $\mathbf{F}$ acts.

Exercise 13.2 *Use this last result to independently confirm Eq. 13.25.*

13.5 Force Transformations

The formulae for transforming forces from one inertial frame to another are readily obtained from the transformations of momentum and time in the definition of force, Eq. 13.1. Thus, in a frame moving with speed V along the positive x axis with respect to a frame in which the force is $\mathbf{F}$, the force will be

$$\mathbf{F}' = \frac{d\mathbf{p}'}{dt'}.$$

Evaluation of this is simplified by application of the chain rule for differentiation:

$$\frac{d\mathbf{p}'}{dt'} = \frac{\frac{d\mathbf{p}'}{dt}}{\frac{dt'}{dt}}. \tag{13.34}$$

Using the Lorentz transformations for time

$$t' = \gamma_V(t - \frac{\beta_V x}{c})$$

and the similar transformations for momentum

$$\begin{aligned} p'_x &= \gamma_V(p_x - \beta_V \frac{E}{c}) \\ p'_y &= p_y \\ p'_z &= p_z \end{aligned}$$

we obtain

$$\begin{aligned} F'_x &= \frac{\gamma_V\left(\frac{dp_x}{dt} - \frac{\beta_V}{c}\frac{dE}{dt}\right)}{\gamma_V\left(1 - \frac{\beta_V}{c}\frac{dx}{dt}\right)} \\ &= \frac{\left(F_x - \beta_V \frac{P}{c}\right)}{\left(1 - \frac{\beta_V}{c}u_x\right)} \end{aligned} \tag{13.35}$$

where $P = \mathbf{F} \cdot \mathbf{u}$ is the power exerted and u_x is the x component of the velocity $\mathbf{u}$ of the particle (on which the force acts) in the old frame.

Exercise 13.3 *Show that*

$$F'_y = \frac{F_y}{\gamma_V\left(1 - \frac{\beta_V u_x}{c}\right)} \tag{13.36}$$

$$F'_z = \frac{F_z}{\gamma_V\left(1 - \frac{\beta_V u_x}{c}\right)}. \tag{13.37}$$

The inverse force transformations are trivially obtained by changing the sign of β_V in the above. Notice that the denominators in these expressions are the same as in the velocity transformations.

It can be seen from the above that force and power mix under transformations. Consider therefore the transformation for power. We have

$$P' = \frac{dE'}{dt'} = \frac{dE'}{dt} \Big/ \frac{dt'}{dt}. \tag{13.38}$$

Since E/c is the time component of 4-momentum, the transformations of this component conveniently give

$$\frac{P'}{c} = \frac{\frac{P}{c} - \beta_V F_x}{\left(1 - \frac{\beta_V u_x}{c}\right)} \tag{13.39}$$

Exercise 13.4 *Verify this.*

13.6 The Minkowski 4-force

The complicated nature of the force transformations arises from the transformation of time in the denominators of the definitions of force and power. A more elegant concept is the Minkowski 4-force, defined by

$$\mathcal{F}^\mu = \frac{dp^\mu}{d\tau}. \tag{13.40}$$

The 4-force may be written as

$$\begin{aligned}\mathcal{F}^\mu &= \gamma_V \frac{dp^\mu}{dt} \\ &= \gamma_V \left(\frac{d(E/c)}{dt}, \frac{d\mathbf{p}}{dt}\right) \\ &= \gamma_V \left(\frac{P}{c}, \mathbf{F}\right). \end{aligned} \tag{13.41}$$

Thus we see that the power, divided by c, is like a time component of force.

It is also useful to write $p^\mu = mu^\mu$ where u^μ is the particle's 4-velocity. Then the 4-force is just

$$\mathcal{F}^\mu = m\frac{du^\mu}{d\tau} = ma^\mu \tag{13.42}$$

where

$$a^\mu \equiv \frac{du^\mu}{d\tau} \tag{13.43}$$

is defined as the 4-acceleration.

13.7 A flexible string paradox

Consider a perfectly flexible, massless string stretched between points A and B. See fig. 13.3. The string makes an angle θ with the x-axis, which we can represent via

$$\tan\theta = L_y/L_x. \tag{13.44}$$

The tension in the string exerts a force on A, directed along the string. In the rest frame of the string, this force has components

$$F_x = T\cos\theta \tag{13.45}$$

$$F_y = T\sin\theta. \tag{13.46}$$

Consider another frame, moving to the right along the x-axis with speed V. In this frame the string is moving to the left and is tilted at a different

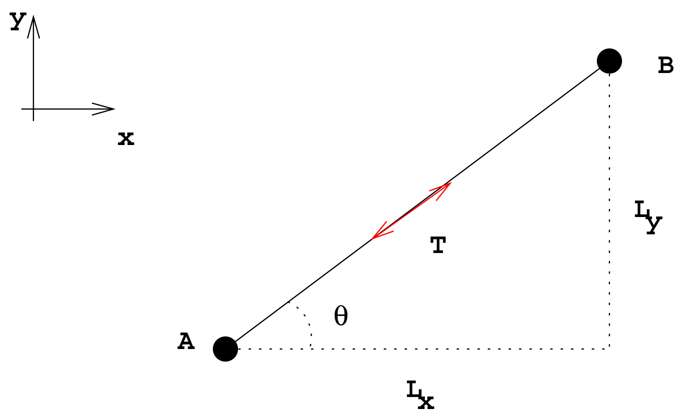

Figure 13.3: A flexible string of length $L = (L_x^2 + L_y^2)^{1/2}$, stretched between A and B in the xy-plane has a tension, T, and makes an angle θ with the x axis.

angle, just as with the tilted stick example of chapter 5. The angle it makes with the x'-axis is given by

$$\tan\theta_s' = L_y'/L_x' = L_y/(L_x/\gamma_V) = \gamma_V \tan\theta. \tag{13.47}$$

The components of force in this new frame are given by the transformation equations.

$$\begin{aligned} F_x' &= \frac{F_x - \beta_V \frac{\mathbf{F}\cdot\mathbf{u}}{c}}{\left(1 - \beta_V \frac{u_x}{c}\right)} \\ F_y' &= \frac{F_y}{\gamma_V \left(1 - \beta_V \frac{u_x}{c}\right)} \end{aligned}$$

where $\mathbf{u}$ is the velocity of A in the old frame. But $\mathbf{u} = 0$. Hence

$$F_x' = F_x = T\cos\theta \tag{13.48}$$
$$F_y' = F_y/\gamma_V = T\sin\theta/\gamma_V. \tag{13.49}$$

However, we observe that the tension is now at an angle given by

$$\tan\theta_t' = F_y'/F_x' = \tan\theta/\gamma_V \neq \tan\theta_s'. \tag{13.50}$$

In this new frame, the tension is not directed along the string. It appears that a perfectly flexible string is exerting a shear force!

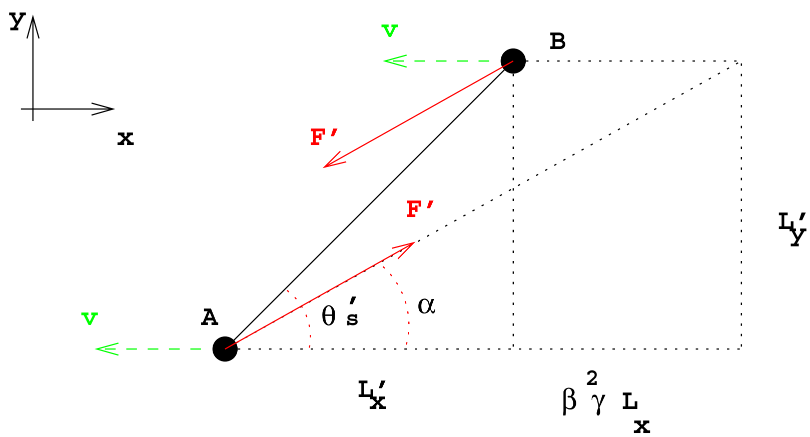

Figure 13.4: The instantaneous position of the string (solid line), in a frame in which it is moving, does not correspond to the transformed coordinates of the instantaneous position of the string in its rest frame. The transformed coordinates make an angle, α, with the x-axis which is precisely that of the transformed force.

To understand what is happening here, one needs to appreciate that the transformed forces, while assumed constant in magnitude in this example, are actually those acting at different times, since what is simultaneous in the string rest frame is not simultaneous in the frame in which it is moving.

Exercise 13.5 *Show that if the forces are evaluated simultaneously in the string rest frame then the transformed forces are those separated in time by* $\Delta t' = \beta\gamma L_x/c$.

It turns out that the direction of the transformed force is precisely in the direction that the other end of the string was at by this amount in the past (or for the other end, where it will be by that amount in the future). See fig. 13.4. To be explicit, the distance travelled by B in time $\Delta t'$ is

$$\Delta x' = v\Delta t' = \beta^2\gamma L_x. \tag{13.51}$$

If we add this to L'_x we get

$$L_x/\gamma + \beta^2\gamma L_x = \gamma(1/\gamma^2 + \beta^2)L_x = \gamma L_x. \tag{13.52}$$

Thus the direction to the other end, B, at $\Delta t'$ in the past is given by

$$\tan\alpha = L'_y/(L_x/\gamma + \beta^2\gamma L_x) = L_y/(\gamma L_x) = \tan\theta/\gamma = \tan\theta'_t \tag{13.53}$$

and so we see that $\alpha = \theta'_t$, i.e. the transformed force on A points towards the location of the other end B at an earlier time while that on B points towards the location of A at a future time. The time difference is just that corresponding to the difference in simultaneity when the string coordinates are transformed directly.

It is of some interest to examine further the deviation of the tension force from the direction of instantaneous orientation of the string. In component form the force on A is

$$\begin{aligned}\mathbf{F}' = (F'_x, F'_y) &= (T\cos\theta, T\sin\theta/\gamma_V) \\ &= (T\cos\theta, \gamma_V T\sin\theta(1-\beta_V^2)\,)\end{aligned} \tag{13.54}$$

Exercise 13.6 *Show that*

$$\cos\theta'_s = \frac{\cos\theta}{\sqrt{\cos^2\theta + \gamma_V^2\sin^2\theta}} \tag{13.55}$$

$$\sin\theta'_s = \frac{\gamma_V\sin\theta}{\sqrt{\cos^2\theta + \gamma_V^2\sin^2\theta}}. \tag{13.56}$$

Using the results from the exercise we see that

$$\begin{aligned}\mathbf{F}' &= (T\cos\theta, \gamma_V T\sin\theta) + (0, -\beta_V^2\gamma_V T\sin\theta) \\ &= (T'\cos\theta'_s, T'\sin\theta'_s) + (0, -\beta_V^2 T'\sin\theta'_s)\end{aligned} \tag{13.57}$$

where $T' = T\sqrt{(\cos^2\theta + \gamma_V^2\sin^2\theta)}$. The first term is in the direction of the instantaneous position of the string. The second term is in the y direction, which is perpendicular to the velocity of A. This second term is also explicitly dependent on the speed of A. It turns out that this second term is closely related to the phenomenon of magnetic forces. (See the next chapter.)

13.8 Force fields

Often, a particle's momentum is altered by other than contact interactions with other bodies. For example, electrically charged particles can repel or attract each other from a distance and massive bodies can similarly attract each other via gravity. Such cases have sometimes been referred to as action-at-a-distance. We know though that particles with a spacelike separation can have no direct influence on each other. Somehow a signal must be propagated, perhaps via exchange of invisible particles, or perhaps there is a breakdown of the symmetry of spacetime in the neighborhood of one of these distant objects that makes it appear as though a force is acting.

Since electrical and gravitational interactions certainly can change the momentum of a particle it is quite meaningful to talk of an electrical or gravitational force. If one considers the work done by these forces, it is found to be independent of path travelled by the particle (while the force acts on it) and dependent only on the endpoints of the path. This is a fact that is by no means true of all forces.

Consider the case where the work done by some force, due to interactions with a distant object, is independent of path. Such forces are called *conservative*. As the force acts on a particle, the energy of that particle changes. It is as if the particle gives up energy to, or takes energy away from, the space through which it moves. Regardless of the detailed mechansim by which this happens (such as collisions with invisible particles) we may think of it in terms of a store of energy in that region of space. The store is like a bank from which energy withdrawals and deposits can be made. We call this store a *field*.

The energy in the field is in general a function of location and time. If the energy at all locations is time independent then the field is called *static*. The energy in the field may be called *potential energy*, and we write it as $U(t, x, y, z)$. Often we simply call U the potential, or the potential field.

Consider a particle moving in the x direction, slowing down as it goes. The particle thus gives up energy to the field and the potential energy stored in the field increases. In a distance dx the increase in stored energy is

$$dU = \frac{\partial U}{\partial x} dx \tag{13.58}$$

where the partial derivative, $\frac{\partial U}{\partial x}$ means, take the change in U as x varies while all other variables (y, z, and t) are held constant. For particle motion in a general direction we have

$$\begin{aligned} dU &= \frac{\partial U}{\partial x}dx + \frac{\partial U}{\partial y}dy + \frac{\partial U}{\partial y}dy \\ &= \left(\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}, \frac{\partial U}{\partial z}\right) \cdot (dx, dy, dz) \\ &\equiv \nabla U \cdot d\mathbf{x}. \end{aligned} \tag{13.59}$$

∇U is known as the gradient of U since it represents the rate of change of U in any direction. By our definition of potential energy, dU is just the loss in kinetic energy of the particle. From the work-energy theorem, this is just the work done on the particle, i.e.

$$dW \equiv \mathbf{F}.d\mathbf{x} = -dU. \tag{13.60}$$

Hence, the force (exerted by the field) acting on the particle may be written as

$$\mathbf{F} = -\nabla U \tag{13.61}$$

where the minus sign indicates that the direction of the force is opposite to the direction of increasing potential. It is to be noted that in determining the force via Eq. 13.61, time is held constant. So it is the instantaneous values of U at $\mathbf{x}$ and $\mathbf{x} + d\mathbf{x}$ that are important.

It is also worth noting that the absolute value of U is not defined, only the gradient of U is a meaningful quantity. Similarly, if we transform to another frame, we see that the field must carry some momentum, but care is required in interpreting directly the space components of the transformed field as momentum.

Review

What equation is used to define force in relativity?

Under what circumstances is the force in the same direction as the acceleration?

What is Newton's third law and explain briefly its physical basis.

How is work defined?

What is the work-energy theorem?

What is power?

In the transformation law for force

$$F'_x = \frac{F_x - \beta_V \frac{\mathbf{F}.\mathbf{u}}{c}}{1 - \beta_V \frac{u_x}{c}}$$

1. What does $\mathbf{u}$ represent?
2. What physical quantity does $\mathbf{F}.\mathbf{u}$ represent?

Write down the transformation law for a component of $\mathbf{F}$ perpendicular to the direction of relative motion of the new frame (which you should take to be to the right along the x-axis).

What is the Minkowski 4-force?

Why is the tension force in a moving string not directed along the string?

The tension force in a string may be resolved into a piece in the direction of the instantaneous position of the string and a remainder. What is the relative direction of the remainder and the velocity of the string?

What is a field?

What is a conservative force?

Questions

1. Is it meaningful, or useful, to apply the concept of force in a non-inertial frame? (Note that in a non-inertial frame a free particle will not travel in a straight line at uniform speed and so the quantity $\mathbf{p} = \gamma m \mathbf{v}$ will appear to change.)

2. Some (conservative) forces, e.g. those due to gravitation and electromagnetism, are very long range. This implies that force fields exist basically everywhere in the universe. Does this negate the validity of the concept of inertial frames?

3. How might one use a well-understood force to determine particle masses?

Problems

1. Show that the general case of Eq. 13.8 can be written as

$$\mathbf{F} = M\mathbf{a} \tag{13.62}$$

where M is a matrix. (M has a very similar form to that of the moment of inertia matrix for rotations of a body about other than one of its principal axes.) Show that M has eigenvalues $\gamma^3 m$ and γm (the latter eigenvalue being repeated, i.e. twofold degenerate).

2. Show, by starting directly from the definition of force, that the magnitude of the centripetal force is

$$F_c = p\omega \tag{13.63}$$

where p is the magnitude of the ordinary momentum of the particle moving in the circle and ω is its angular velocity.

3. Consider the flexible string paradox.

   (a) Evaluate the explicit connection between force and acceleration, Eq. 13.8, in the frame in which the string is moving. Using the components of this equation, show that the acceleration, $\mathbf{a}'$, of either end of the string, in this frame, is in the direction of the instantaneous orientation of the string.

   (b) Verify this by directly determining (and evaluating) the transformations for acceleration. Show that the result is true not only for the initial configuration where the particles at either end of the string are stationary (in the string frame) but that it remains true after the particles have begun to move.

   (c) Explain the physical basis of this result by considering the definition of acceleration and the changing positions of the particles.

4. (a) Show that the rate at which a force does work on an object is

$$P = \gamma^3 m\mathbf{a} \cdot \mathbf{v} \tag{13.64}$$

   and hence: if $\mathbf{a} \perp \mathbf{v}$ the force does no work.

(b) Show that

$$\mathbf{F} = \gamma m\mathbf{a} + \frac{P}{c}\beta \tag{13.65}$$

and also that

$$\frac{d\gamma}{dt} = \frac{P}{E_0}. \tag{13.66}$$

5. (a) Show that the (passive) transformation equations for force can be rewritten as

$$F'_x = F_x - \beta_V\gamma_V(F_y\frac{u'_y}{c} + F_z\frac{u'_z}{c}) \tag{13.67}$$

$$F'_y = F_y\frac{u'_y}{u_y} = \gamma_V F_y(1 + \beta_V\frac{u'_x}{c}) \tag{13.68}$$

$$F'_z = F_z\frac{u'_z}{u_z} = \gamma_V F_z(1 + \beta_V\frac{u'_x}{c}). \tag{13.69}$$

Hence conclude that if $\mathbf{F}$ is velocity independent then the corresponding velocity independent piece in the new frame is

$$\mathbf{F}'_{\text{vel. ind.}} = (F_x, \gamma_V F_y, \gamma_V F_z). \tag{13.70}$$

(b) Show that if a force $\mathbf{F}$ arises from a static field surrounding a fixed source (and is accordingly velocity independent) and is such that it is always directed towards the source, then in any other frame the velocity independent piece of the transformed force $\mathbf{F}'$ will always be directed towards the instantaneous position of the source.

6. (a) Show that the acceleration, $\mathbf{a}$, of a particle with velocity $\mathbf{v}$ under the influence of a force, $\mathbf{F}$, is

$$\mathbf{a} = \frac{1}{\gamma m}\left[\mathbf{F} - \frac{\mathbf{F}\cdot\mathbf{v}}{c^2}\mathbf{v}\right]. \tag{13.71}$$

(b) Show that the components parallel and perpendicular to $\mathbf{F}$ are

$$a_{\parallel} = \frac{F}{\gamma m}\left[1 - \beta^2\cos^2\theta_{vF}\right] \tag{13.72}$$

$$a_{\perp} = \frac{F}{\gamma m}\sqrt{\beta^2(\beta^2-1)\cos^2\theta_{vF}} \tag{13.73}$$

and that **a** has magnitude

$$a = \frac{F}{\gamma m}\sqrt{1+\beta^2(\beta^2-2)\cos^2\theta_{vF}} \tag{13.74}$$

where θ_{vF} is the angle between **v** and **F**.

(c) Show that the angle, θ_{aF}, between **a** and **F** is given by

$$\cos\theta_{aF} = \frac{1-\beta^2\cos^2\theta_{vF}}{\sqrt{1+\beta^2(\beta^2-2)\cos^2\theta_{vF}}} > 0 \tag{13.75}$$

and verify that **a** and **F** are parallel for $\beta = 0$ or $\cos\theta_{vF} = 0,\ 1$. Further show that $\cos\theta_{aF}$ assumes its minimum value

$$(\cos\theta_{aF})_{\min} = \frac{2\sqrt{1-\beta^2}}{2-\beta^2} = \frac{2\gamma}{\gamma^2+1}$$

(which defines the *maximum* of θ_{aF}) for

$$\cos\theta_{vF} = \pm\frac{1}{\sqrt{2-\beta^2}}.$$

Observe that though $\cos\theta_{aF}$ tends to zero at large particle speeds (implying that **a** and **F** are perpendicular) that in the ultrarelativistic limit the magnitude of the acceleration tends to zero.

(d) Observe that Eq. 13.75 restricts θ_{aF} to between $-90°$ and $+90°$ but does not suffice to determine the direction of **a**. To settle this, show that the angle, $\cos\theta_{av}$, between **a** and **v** is given by

$$\cos\theta_{av} = \frac{F}{\gamma^3 ma}\cos\theta_{vF} = \frac{\cos\theta_{vF}}{\gamma^2\sqrt{1+\beta^2(\beta^2-2)\cos^2\theta_{vF}}}.$$

Prove that

$$|\cos\theta_{aF}| > |\cos\theta_{av}|$$

and thus deduce the direction of **a**.

Chapter 14

Electricity and Magnetism

Electrical phenomena are common and can even be said to be responsible for many aspects of the world as we know it. Perhaps the simplest manifestation of electricity occurs when certain objects are rubbed and thereby become "charged." The Ancient Greeks knew that if amber was rubbed with fur then the fur crackled and the hairs stood on end, while the amber could attract small objects such as hair. Indeed the word electricity comes from the Greek word, electrum, for amber. By the eighteenth century it was understood that electric forces could be either repulsive or attractive and we now understand this to be due to two different kinds of *electric charge*, which we call positive and negative, following the terminology of the eighteenth century American scientist Benjamin Franklin.

Magnetic phenomena have also been known for thousands of years. Pieces of the natural mineral magnetite, called lodestones, were known to exert repulsive and attractive forces on each other — the force changing as the lodestone was turned around. Freely suspended lodestones were also known to orient themselves in a north-south direction and were the precursor of the modern compass. Such objects we now call magnets and understand that the behavior of the compass is due to Earth itself being a magnet.

It is our goal in this chapter to show that magnetism is just a manifestation of electricity in inertial frames in which electric charges are moving.

14.1 Historical note

Until the nineteenth century, electricity and magnetism were largely curiosities. In 1820, Ampère and Oersted showed that magnetic effects could be produced by electric charges that moved, i.e. electric currents, and the connection between the two was further elucidated by Faraday, who also showed that electric currents could be induced by moving magnets. This work led to the electric motor and paved the way for modern technology.

In the late 1860s, Maxwell gathered together the known results from experiments with electricity and magnetism and formulated them into a unified mathematical theory based on four equations known as Maxwell's equations. The feat in itself was brilliant but was made even more stunning when Maxwell discovered that his equations indicated that oscillating electric and magnetic fields could propagate as a wave whose speed in vacuum was coincidentally that of light. It was thus hypothesized that light was an electromagnetic wave.

Much interest in Maxwell's equations led to the discovery that his theory was invariant under Lorentz transformations. Indeed, these transformations are named after their discoverer, who found them before Einstein published his relativity theory. However, the significance of the transformations was not understood. Scientists of those times assumed that electromagnetic waves needed some sort of medium for their propagation. They termed it the luminiferous ether. The classic experiment of Michelson and Morley was an attempt to measure the speed of the Earth through the ether.

Although Einstein was aware that attempts to measure the speed of the "ether wind" had all failed, this played little role in his development of relativity theory. He was impressed by the beauty of Maxwell's equations and believed that one should accept they were the same in all inertial reference frames. (Other physicists had tried to modify them to take account of the speed of the observer with respect to the ether.) He was also struck by the symmetry between inducing a current in a loop of wire by moving a magnet through it and by doing the same by moving the wire loop past a stationary magnet. This can be explained with Maxwell's theory in two ways. Firstly, one can say that a changing magnetic field inside the loop induces an electric field in the wire which causes an electric current. (This is Faraday induction.) Secondly, one can say, in the case of the moving loop only, that charges in the wire are moving in a magnetic field and thus experience a magnetic force — that being velocity dependent — thereby causing free charges in the wire to flow as a current. Einstein was convinced that only relative motion was important and thus the physics involved was identical. The two experiments should be regarded as one and the same but observed in different inertial frames.

Thus electromagnetic phenomena and Maxwell's theory were instrumental in the historical development of relativity. This historical development is largely responsible for the common version of Einstein's Principle of Relativity: "The laws of physics are the same in all inertial frames." In the development presented here though, we have no prior notion of what a "law of physics" is and indeed defined it in terms of being the same in all inertial frames.

Accordingly, in this chapter we begin with only a rudimentary knowledge of electromagnetic phenomena and seek to apply what we have learned about the consequences of Poincaré symmetry.

14.2 Electric charge

The two types of electric charge can be considered as opposites and thus the terminology of positive and negative is very appropriate. We usually represent them by $+Q$ and $-Q$, or similar notation. The generation of electrical effects by rubbing amber with fur, or walking across a carpet in rubber shoes, etc., can be considered as a separation of the two types of charge. Normally, both are present in a material in equal amounts and so exactly cancel each other out. The material is then said to be electrically neutral (the positive and negative charges summing to zero) and no electric forces are observed.

When the charges are separated, like charges repel each other and unlike charges attract. The electric force of repulsion or attraction depends on the amount of charge present. These are experimental facts.

Nowadays we know that electric charge is quantized. The smallest observable unit of free charge is that carried by the electron and is classed as negative. A positive charge of the same magnitude is carried by the proton. It is often called, $+e$. Although there is strong evidence that the proton is comprised of constituent particles, called quarks, with charges that are $\frac{2}{3}e$ and $-\frac{1}{3}e$, these quarks appear unable to exist outside of the proton. Thus our statement that the proton charge is the smallest observable free charge remains true.

All experiments are consistent with the electric charge being a Poincaré invariant quantity. At the elementary particle level it is a characteristic property of the particle. Furthermore, in any physical process, charge is conserved. Charges may be transferred from one particle to another but they never disappear — except in oppositely charged pairs, so that the net charge is conserved.

14.3 Electrostatic Potential

The experimental evidence is that the electric force exerted by a stationary electric charge is conservative. Thus it may be represented as the gradient of a potential, as discussed in the last chapter. It is also found though that the electric force on a charged object is proportional to the charge carried by that object. Thus it is convenient to define the electrostatic potential, V, as the potential energy, U per unit charge. To be precise, we note that two charged objects will act on each other (according to Newton's Third Law) and so will influence the field configuration of the other, the resultant field being a property of the pair. To be able to speak of the field due to a charge we take a very small test charge, q and define the *electrostatic potential* as

$$V \equiv \lim_{q \to 0} \frac{U}{q}. \tag{14.1}$$

The electric force on this charge is thus

$$\mathbf{F} = q\mathbf{E} \tag{14.2}$$

where

$$\mathbf{E} \equiv -\nabla V \tag{14.3}$$

is the *electric field*.

Of course, the limit in Eq. 14.1 does not exist in a physical sense because of charge quantization. Nevertheless we can define the limit mathematically and treat the fields in this classical approximation. Then the field is a property only of its *source* charge. In particular, the electric field (and electric force) is independent of the velocity of the test charge. We can take this velocity independence of Eq. 14.2 as a general characteristic of an electric field that we would like to preserve in any frame.

Since the electric force is proportional to the (test) charge q, and the electric field is a function solely of the source, Eq. 14.2 must be true for any charge q. In addition, the electric force due to different sources must add vectorially (as do all forces) and so therefore do the electric fields. It follows that the electric potentials of different sources just add.

It is important to understand that Eq. 14.3, for the electric field, was written down for a stationary source. In general, a more complicated expression could be admitted.

Exercise 14.1 *Use Eq. 14.3 to show that*

$$\int_A^B \mathbf{E} \cdot d\mathbf{x} = -(V_B - V_A).$$

(Note that this is a line integral as in Eq. 13.19.) Thus, if one integrates around an arbitrary closed loop the corresponding (closed) line integral is

$$\oint \mathbf{E} \cdot d\mathbf{x} = 0. \tag{14.4}$$

(This is a special case of one of Maxwell's equations.)

Physically, Eq. 14.4 is just a reflection of the fact that the electric force, in this frame, is conservative.

14.4 Gauss's law

In 1755 Benjamin Franklin observed that a charged cork ball placed inside a charged metal container is not attracted to the walls of the container. He reported this to Priestley who noted a similarity with Newton's theory of gravitation and concluded that the force laws must be similar. The absence of an electric field inside a charged container was recognized by Henry Cavendish in 1773 as a consequence of Gauss's law. For the electric field this law can be expressed as

$$\oint \mathbf{E} \cdot d\mathbf{A} = k_E Q_{\text{enc}}. \tag{14.5}$$

The integral here is one over a closed surface of the component of $\mathbf{E}$ perpendicular to the surface — the area element, d$\mathbf{A}$ being a vector directed perpendicular to the surface and outward. One calls the quantity

$$d\Phi_E = \mathbf{E} \cdot d\mathbf{A} \tag{14.6}$$

the *electric flux*. Q_{enc} is the charge enclosed by the surface and the proportionality constant, k_E depends on the system of units used. Common choices are

$$k_E = \begin{cases} 4\pi & \text{Gaussian (cgs) units} \\ 1/\varepsilon & \text{SI units} \end{cases} \tag{14.7}$$

where ε is the electric permittivity in the region of the electric field. We shall always take that region to be a vacuum, for which one uses the notation, ε_0.

Gauss's law and Eq. 14.4 are sufficient to determine the electric field about any stationary set of charges. Sometimes though a short cut is possible by applying Gauss's law and any symmetry of the charge configuration.

14.4.1 Coulomb's law: field about a stationary point charge

Consider the electric field about a stationary point charge, +Q. We know that a positive test charge will be repelled by the charge Q and so $\mathbf{E}$ must point away from it. By the spherical symmetry of the situation one can conclude that $\mathbf{E}$ must be directed radially outwards and its magnitude E depends only on the distance from Q and not at all on the orientation. Gauss's law applies to an arbitrary closed surface but we can exploit the spherical symmetry by using a spherical (Gaussian) surface, of radius r, centered on Q. See fig. 14.1. Since $\mathbf{E}$ and $d\mathbf{A}$ are both perpendicular to this Gaussian surface and since E is uniform over it we have

$$\mathbf{E} \cdot d\mathbf{A} = E\, dA$$

where dA is the magnitude of the area element. Furthermore, since E is constant over the surface it can be extracted from the integral. Thus

$$\oint \mathbf{E} \cdot d\mathbf{A} = E \oint dA = E(\text{total area}) = E(4\pi r^2). \tag{14.8}$$

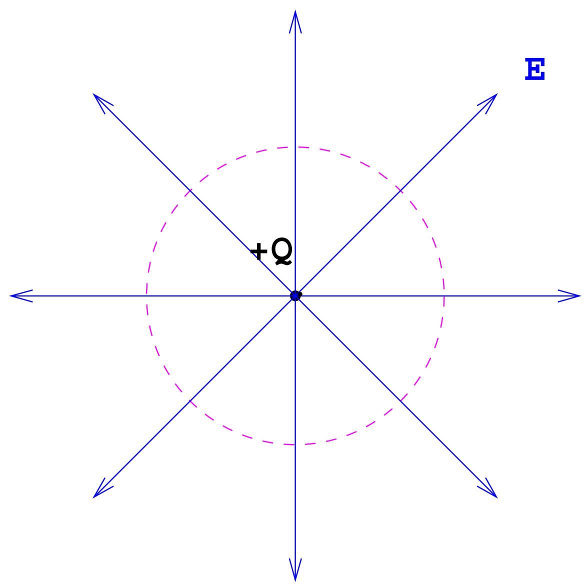

Figure 14.1: Electric field and Gaussian surface about a point charge.

The charge enclosed by this surface is just Q and so Gauss's law implies that

$$E = k_C \frac{Q}{r^2} \tag{14.9}$$

where

$$k_C = \frac{k_E}{4\pi} \tag{14.10}$$

is called Coulomb's constant. We may infer that a possible choice for the electrostatic potential is

$$V = -k_C \frac{Q}{r} \tag{14.11}$$

and that the electric force is

$$F = k_C \frac{Qq}{r^2}. \tag{14.12}$$

Eq. 14.12 is known as Coulomb's law. The r^{-2} dependence is particularly noteworthy. It follows from the spherical geometry and Gauss's law. A force law with this distance behavior is said to be an *inverse-square* law. Deviations from inverse-square behavior would therefore signal a breakdown of Gauss's law. If we write the force as

$$F \propto \frac{1}{r^{2\pm\delta}} \tag{14.13}$$

then current experimental limits show that [25]

$$\delta < 3 \times 10^{-16}. \tag{14.14}$$

Exercise 14.2 *Consider the closed path formed by starting at the point charge Q in fig. 14.1, following a field line outwards to the spherical Gaussian surface, traversing the Gaussian surface and then following another field line back to the point charge. For this closed path and the electric field of Eq. 14.9, show that Eq. 14.4 holds true.*

14.4.2 Field about a stationary parallel plate capacitor

Consider two plane sheets separated by a uniform distance d, one sheet carrying charge $+Q$ and the other $-Q$. Such an arrangement is called a parallel plate capacitor. For simplicity, let d be small compared to the size of the sheets so that we can approximate the sheets as extending to infinity in all directions. We now apply Gauss's law to determine the electric field. $\mathbf{E}$ must point away from the positively charged sheet and towards the negatively charged sheet. By symmetry we expect the field to be perpendicular to each sheet and to be uniform in strength at a fixed distance from a sheet (there being no preferred part of the sheet). See fig. 14.2. Let the charge per unit area be called σ. For both Gaussian surfaces in fig. 14.2, the only contributions to the closed surface integral of the electric flux are those from surfaces where $\mathbf{E}$ is perpendicular to the surface and thus

$$\mathbf{E} \cdot d\mathbf{A} = E\, dA.$$

Consider first the Gaussian surface that encloses just the positive sheet and ignore for the moment the presence of the other. By symmetry, the strength, E of the electric field should be the same on both sides. Accordingly,

$$\oint \mathbf{E} \cdot d\mathbf{A} = E \oint dA = E(2A) \tag{14.15}$$

where A is the surface area of one parallel face of the Gaussian surface. The charge enclosed by this surface is

$$Q_{\text{enc}} = \sigma A \tag{14.16}$$

and so

$$E = k_E \sigma / 2 \tag{14.17}$$

independent of distance from the sheet.

Similarly, one finds for the negative sheet that

$$E = k_E \sigma / 2,$$

which is of the same strength, but points towards the negative sheet. Combining the two results, we just add the two contributions together (vectorially) and find

$$E = \begin{cases} k_E \sigma & \text{inside the capacitor} \\ 0 & \text{outside the capacitor} \end{cases} \tag{14.18}$$

The result for the region outside of the capacitor is consistent with Gauss's law for the Gaussian surface enclosing both sheets but does not follow directly from it because sufficient information on the direction of $\mathbf{E}$, in the exterior region, is lacking.

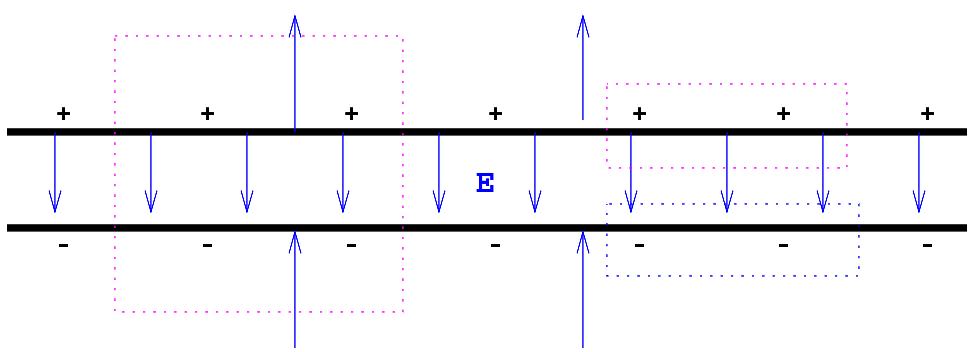

Figure 14.2: A side-on view of the electric field for an infinite parallel plate capacitor. Three box-like Gaussian surfaces are shown, one enclosing both sheets and the others enclosing just one sheet. The field outside turns out to be zero.

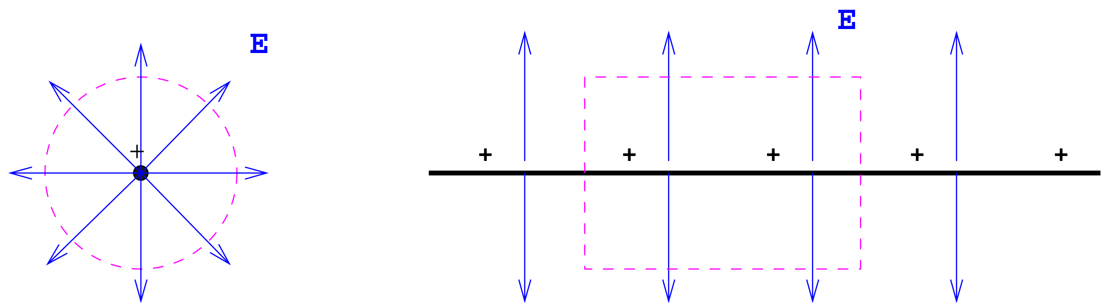

Figure 14.3: End-on and side-on view of the electric field about a long line of charge, with a cylindrical Gaussian surface indicated.

Note that the magnitude of E is the same at all points inside the capacitor. The electric potential therefore varies linearly from an arbitrary value, which we can take as zero, on the negatively charged plate and $k_E \sigma d$ on the positive plate.

14.4.3 Field about a stationary line of charge

Consider a long line of charge. Let the charge density per unit length be λ. By symmetry, the electric field must point radially away from the line of charge and be uniform along its length. See fig. 14.3. The cylindrical symmetry of the configuration may be exploited using a cylindrical Gaussian surface centered on the line of charge.

The only contribution to the Gaussian integral is over the curved surface of the cylinder. If the cylinder's length is l and its radius r, then the area of this curved surface is $2\pi rl$. Thus the Gaussian integral is

$$\oint \mathbf{E} \cdot d\mathbf{A} = E(2\pi rl). \tag{14.19}$$

Since the charge enclosed by the surface is λl, it follows by Gauss's law that

$$E = \frac{k_E}{2\pi}\frac{\lambda}{r}. \tag{14.20}$$

14.5 Transformed electric field

The components of the electric field about a stationary charge are given by

$$E_x = -\tfrac{\partial V}{\partial x} = \lim_{\Delta x \to 0} -\frac{\Delta_x V}{\Delta x} \tag{14.21}$$

$$E_y = -\tfrac{\partial V}{\partial y} = \lim_{\Delta y \to 0} -\frac{\Delta_y V}{\Delta y} \tag{14.22}$$

$$E_z = -\tfrac{\partial V}{\partial z} = \lim_{\Delta z \to 0} -\frac{\Delta_z V}{\Delta z} \tag{14.23}$$

where

$$\Delta_x V(x,y,z,t) \equiv V(x+\Delta x,y,z,t) - V(x,y,z,t) \tag{14.24}$$

$$\Delta_y V(x,y,z,t) \equiv V(x,y+\Delta y,z,t) - V(x,y,z,t) \tag{14.25}$$

$$\Delta_z V(x,y,z,t) \equiv V(x,y,z+\Delta z,t) - V(x,y,z,t) \tag{14.26}$$

respectively.

We now wish to consider what we mean by an electric field in another frame. Let us start in a frame in which the source is at rest and thus the field has only an energy piece and does not carry momentum. In this frame, V is static, i.e. it does not depend on time. Since electric potential is just energy per unit charge (and charge is invariant), it must transform in the same way as energy. Hence in a new frame we must have

$$V'(x',y',z',t') = \gamma V(x,y,z,t) \tag{14.27}$$

where x', y', z' and t' are the transformed coordinates of x, y, z and t. But because of the space separation in $\Delta_x V$ the corresponding $\Delta_{x'} V'$ will involve the potential at different times. Furthermore, if we consider the inverse transformation, we see that $\Delta_x V$ must also involve the spatial part of an energy-momentum 4-vector in the new frame. It seems that Eq. 14.3 is not a satisfactory basis for defining the electric field in the new frame.

To clarify the situation, consider the electric field between a parallel plate capacitor. When the capacitor is stationary with respect to the observer, the field is uniform in between the plates and directed perpendicularly to them. (Note that the field is also directed along the line between charges opposite each other.) Thus a charged particle entering the region in between the plates will experience an electric force directed along the electric field lines

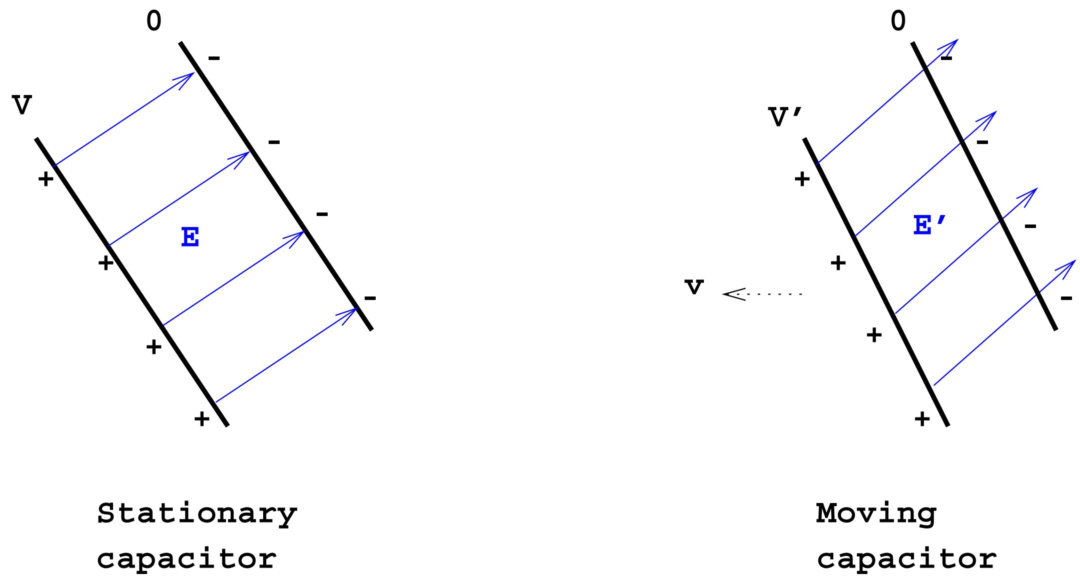

Figure 14.4: A parallel plate capacitor at rest and moving. (The length of the electric field vectors is not related to the scale of the capacitor.)

and *independent of the particle's velocity*. Now consider the same capacitor in a frame moving along the positive x axis. In this new frame the capacitor is moving in the negative x direction and will appear tilted with respect to its rest frame orientation. See fig. 14.4. The forces acting on a positive charge between the plates are oriented as in the flexible string paradox, discussed in the chapter 13. We identify the electric field in the new frame as arising from the velocity independent piece of this transformed force.

To isolate the velocity independent piece of the force consider a test charge with arbitrary velocity $\mathbf{u}$ in the capacitor rest frame. The total force is given by Eq. 14.2, where $\mathbf{E} = (E_x, E_y, E_z)$. In the new frame the velocity of the test charge has components

$$\begin{aligned} u'_x &= \frac{u_x - \beta c}{(1 - \beta \frac{u_x}{c})} \\ u'_y &= \frac{u_y}{\gamma(1 - \beta \frac{u_x}{c})} \\ u'_z &= \frac{u_z}{\gamma(1 - \beta \frac{u_x}{c})}. \end{aligned}$$

whereas the force has components

$$\begin{aligned} F'_x &= q\frac{E_x - \beta(E_x u_x + E_y u_y + E_z u_z)/c}{(1-\beta\frac{u_x}{c})} \\ &= qE_x - q\beta\gamma E_y \frac{u'_y}{c} - q\beta\gamma E_z \frac{u'_z}{c} && (14.28) \\ F'_y &= q\frac{E_y}{\gamma(1-\beta\frac{u_x}{c})} = q\frac{\gamma E_y}{(1-\beta\frac{u_x}{c})}(1-\beta^2) \\ &= q\gamma E_y(1+\beta u'_x/c) && (14.29) \\ F'_z &= q\gamma E_z(1+\beta u'_x/c). && (14.30) \end{aligned}$$

Thus we see that $\mathbf{F}'$ has two terms: one independent of $\mathbf{u}'$ and one that depends on $\mathbf{u}'$. The term which is independent of $\mathbf{u}'$ looks like an electric force $\mathbf{F}' = q\mathbf{E}'$ where

$$\mathbf{E}' = (E_x, \gamma E_y, \gamma E_z). \qquad (14.31)$$

This justifies our identification of $\mathbf{E}'$ as given by the velocity independent piece.

Looking at $\mathbf{E}'$ component by component:

$$\begin{aligned} E'_x &= E_x && (14.32) \\ E'_y &= \gamma E_y && (14.33) \\ E'_z &= \gamma E_z. && (14.34) \end{aligned}$$

We see that the component of electric field in the direction of motion is unaltered while the components perpendicular to the motion are increased. Thus, if we consider the angle that $(E_x, \gamma E_y, 0)$ makes with the x-axis we find

$$\tan\theta' = \frac{E'_y}{E'_x} = \gamma\frac{E_y}{E_x} = \gamma\tan\theta, \qquad (14.35)$$

where θ is the angle in the original frame. A similar result is obviously obtained for fields in the xz plane. We observe that the original static $\mathbf{E}$ field has been tilted to $\mathbf{E}'$ exactly as in the tilted stick example. Since the original field was static and directed towards the source charges it follows that $\mathbf{E}'$ is directed towards the instantaneous position of the sources in the new frame.

14.5.1 Frame independence of Gauss's law

One striking consequence of the transformations for $\mathbf{E}$ is that the electric flux is the same in any inertial frame. Recalling that $d\mathbf{A}$ represents a vector perpendicular to the surface with area element of magnitude dA we see from Lorentz contraction that

$$\begin{aligned} dA'_x &= dy'dz' &= dA_x \quad (14.36) \\ dA'_y &= dz'dx' &= dA_y/\gamma \quad (14.37) \\ dA'_z &= dx'dy' &= dA_z/\gamma. \quad (14.38) \end{aligned}$$

Thus

$$\mathbf{E}' \cdot d\mathbf{A}' = E_x dA_x + \gamma E_y \frac{dA_y}{\gamma} + \gamma E_z \frac{dA_z}{\gamma} = \mathbf{E} \cdot d\mathbf{A}. \quad (14.39)$$

It follows that Gauss's law for the electric flux is valid in any inertial frame. It may therefore be considered as a law of physics.

14.5.2 Electric field of moving point charge

Consider the electric field about a point charge, Eq. 14.9. The full vector field is

$$\mathbf{E} = k_C \frac{Q}{r^2} \mathbf{e}_r \quad (14.40)$$

where

$$\mathbf{e}_r \equiv \frac{\mathbf{r}}{r} \quad (14.41)$$

is a unit vector directed radially outward.

Now consider a frame moving in the positive x direction. The electric field in this new frame is an instantanteous configuration so the relevant distance $\mathbf{r}'$ from the charge is a proper length. Thus

$$\begin{aligned} x &= \gamma x' \\ y &= y' \\ z &= z'. \quad (14.42) \end{aligned}$$

For simplicity, we shall consider just the field in the xy-plane. Thus $\mathbf{r} = (x, y, 0)$ and

$$\begin{aligned} E_x &= k_C Q \frac{x}{(x^2+y^2)^{3/2}} \\ E_y &= k_C Q \frac{y}{(x^2+y^2)^{3/2}}. \end{aligned} \tag{14.43}$$

Applying the transformation rule for the electric field components, and using Eq. 14.42 in Eq. 14.43, we obtain

$$\mathbf{E}' = (E_x, \gamma E_y) = \frac{k_C Q}{[(\gamma x')^2 + y'^2]^{3/2}} \gamma (x', y') \tag{14.44}$$

showing that $\mathbf{E}'$ is in the same direction as $\mathbf{r}' = (x', y')$, as expected.

The magnitude of the transformed field is

$$E' = (E'^2_x + E'^2_y)^{1/2} \tag{14.45}$$

$$= \frac{\gamma k_C Q}{[(\gamma x')^2 + y'^2]^{3/2}} (x'^2 + y'^2)^{1/2} \tag{14.46}$$

$$= \frac{\gamma k_C Q r'}{\gamma^3 [x'^2 + (y'/\gamma)^2]^{3/2}} \tag{14.47}$$

$$= \frac{k_C Q r'}{\gamma^2 [x'^2 + y'^2(1-\beta^2)]^{3/2}} \tag{14.48}$$

$$= \frac{k_C Q r'}{\gamma^2 r'^3 \left[1 - \beta^2 \frac{y'^2}{x'^2 + y'^2}\right]^{3/2}}. \tag{14.49}$$

But

$$\frac{y'}{(x'^2 + y'^2)^{1/2}} = \sin\phi' \tag{14.50}$$

where ϕ' is the angle between the electric field and the x'-axis. Hence,

$$\mathbf{E}' = \frac{k_C Q}{r'^2} \frac{1-\beta^2}{(1-\beta^2 \sin^2\phi')^{3/2}} \mathbf{e}_{r'}. \tag{14.51}$$

This field is weaker in the direction of motion than perpendicular to it. This is shown schematically in fig. 14.5. Note that the electric field in the new

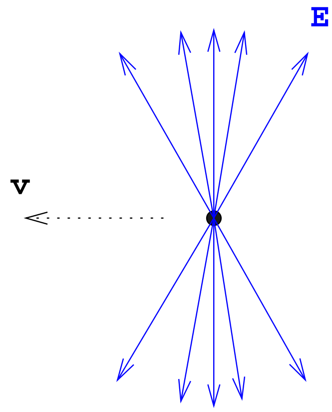

Figure 14.5: The electric field of a moving charge is concentrated in the directions perpendicular to the direction of motion.

frame is symmetric in the forwards and backwards directions. This follows immediately from the behavior of the sine function in Eq. 14.51 and its appearance there in squared form. Physically though, the symmetry follows from the same behavior of the tilted stick and the instantaneous nature of **E**.

Exercise 14.3 *Show that for this field, $\oint \mathbf{E}' \cdot d\mathbf{x}' \neq 0$. (Use the same approach as suggested in the corresponding exercise for the static field about a point charge.) Thus this field is not the gradient of a potential. It is not a field that any static charge configuration could possess.*

14.6 Magnetic force

Consider the transformed force for a general electric field, $\mathbf{E} = (E_x, E_y, E_z)$. The transformations from a frame in which the charges are at rest, i.e. Eqs. 14.28-14.30, can be rewritten as

$$F'_x = qE'_x - q\beta(E'_y \frac{u'_y}{c} + E'_z \frac{u'_z}{c}) \tag{14.52}$$

$$F'_y = qE'_y + q\beta E'_y \frac{u'_x}{c} \tag{14.53}$$

$$F'_z = qE'_z + q\beta E'_z \frac{u'_x}{c}. \tag{14.54}$$

The additional terms, depending on the velocity $\mathbf{u}'$ of the test charge, constitute what is called the magnetic force. It is convenient to rewrite this contribution by noting that it is of the form

$$\mathbf{F}'_m = q\frac{\mathbf{u}'}{k_B c} \times \mathbf{B}' \tag{14.55}$$

where

$$\mathbf{B}' \equiv k_B \frac{\mathbf{v_s}'}{c} \times \mathbf{E}' \tag{14.56}$$

is defined as the *magnetic field* and $\mathbf{v_s}'$ is the velocity of the source of $\mathbf{E}'$. (In the above, $\frac{\mathbf{v_s}'}{c} = (-\beta, 0, 0)$.) The proportionality constant k_B depends on the system of units used. Common values are:

$$k_B = \begin{cases} 1 & \text{Gaussian (cgs) units} \\ 1/c & \text{SI units} \end{cases} \tag{14.57}$$

Note that in the cgs system $\mathbf{B}$ and $\mathbf{E}$ have the same units.

The total force on the test charge, moving with velocity $\mathbf{u}'$, is thus

$$\mathbf{F}' = q\mathbf{E}' + q\frac{\mathbf{u}'}{k_B c} \times \mathbf{B}' \tag{14.58}$$

and is valid in any frame. This equation is known as the *Lorentz force law.*

One may take the Lorentz force in any frame and transform it to any other, moving with velocity $\mathbf{v}$ with respect to the first, showing that the general (passive) transformations of electric and magnetic fields are

$$\begin{aligned} E'_{\|} &= E_{\|} \\ B'_{\|} &= B_{\|} \\ \mathbf{E}'_{\perp} &= \gamma(\mathbf{E}_{\perp} + \frac{1}{k_B c}\mathbf{v} \times \mathbf{B}) \\ \mathbf{B}'_{\perp} &= \gamma(\mathbf{B}_{\perp} - \frac{k_B}{c}\mathbf{v} \times \mathbf{E}). \end{aligned} \tag{14.59}$$

The proof of this is somewhat tedious and is left as an exercise. Note that these transformations are rather different from those of 4-vectors. The components of $\mathbf{E}$ and $\mathbf{B}$ actually constitute six independent (non-zero) components of a 16 component object that mathematicians call a tensor. Tensors provide an elegant description of the transformations but there is no need for us to indulge them further here.

14.7 Electric Current

A useful concept when discussing moving source charges is that of an *electric current*. This is defined as

$$i = \frac{dQ}{dt}. \tag{14.60}$$

The meaning of this equation is that the current, i, is the amount of charge, dQ, crossing a plane in time dt. If the charge is moving with speed v then in time dt it travels a distance dx. The charge dQ thus occupies a volume whose length is dx and dQ/dx is the linear charge density, λ. Thus

$$i = \frac{dQ}{dx}\frac{dx}{dt} = \lambda v \tag{14.61}$$

14.7.1 Current density

It is sometimes useful to employ the *current density*, $\mathbf{j}$, which is a vector pointing in the direction of current flow with magnitude

$$j = \frac{i}{A} \tag{14.62}$$

where A is an area element perpendicular to the direction of current flow. Thus, for arbitrary surfaces, S, the current flowing through S is

$$i = \int_S \mathbf{j} \cdot d\mathbf{A}. \tag{14.63}$$

If we also introduce the charge density,

$$\rho = \frac{Q}{V}, \tag{14.64}$$

where V is a volume over which we are averaging, then we find that

$$\mathbf{j} = \frac{\lambda \mathbf{v}}{A} = \rho \mathbf{v}. \tag{14.65}$$

If the charge density is ρ_0 in a frame in which the charges are at rest, then, in any other frame,

$$\rho = \frac{Q}{Al} = \gamma\frac{Q}{Al_0} = \gamma\frac{Q}{V_0} = \gamma\rho_0. \tag{14.66}$$

Therefore

$$\mathbf{j} = \rho_0 \gamma \mathbf{v}. \tag{14.67}$$

Thus one notes that the combination

$$j^\mu = (\rho c, \mathbf{j}) \tag{14.68}$$

can be written as

$$j^\mu = \rho_0 \gamma (c, \mathbf{v}) = \rho_0 u^\mu \tag{14.69}$$

where u^μ is the 4-velocity of the charges. Hence j^μ transforms as a 4-vector in the same way as u^μ.

14.8 Magnetic field surrounding a line of moving charge

If we are in a frame in which a line of charge is at rest then there is an electric field radiating outwards from the line of charge, whose magnitude we found earlier from application of Gauss's law. See Eq. 14.20.

In a frame, moving to the right with speed v along the line of charge, we can again apply Gauss's law to determine the electric field. The derivation is exactly as before and we obtain

$$E' = k_E \frac{\lambda'}{2\pi r'} \tag{14.70}$$

where λ' is the linear charge density in the new frame and the distance r' from the line of charge is perpendicular to the direction of relative motion of the frames and so is the same as the distance r in the original frame. The linear charge density is the charge per unit length,

$$\lambda' = \frac{Q}{l'}. \tag{14.71}$$

Since the length, l' is contracted compared to the original frame we find

$$\lambda' = \gamma \frac{Q}{l} = \gamma \lambda. \tag{14.72}$$

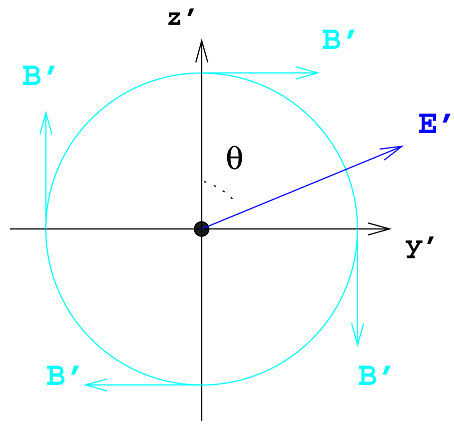

Figure 14.6: Magnetic field about a line of moving charge. The current is into the page and **B** is directed clockwise in uniform circles about the current.

Thus

$$E' = \gamma k_E \frac{\lambda}{2\pi r} = \gamma E \tag{14.73}$$

as expected for an electric field perpendicular to the direction of relative motion of the two frames.

In the new frame, the moving charges generate a magnetic field given by Eq. 14.56. The velocity of the sources is

$$\mathbf{v}' = (-v, 0, 0) \tag{14.74}$$

and the electric field in this frame is the vector

$$\mathbf{E}' = (0, E' \sin\theta, E' \cos\theta) \tag{14.75}$$

where θ is the angle between $\mathbf{E}'$ and the z axis. See fig. 14.6. The vector product is

$$\mathbf{v}' \times \mathbf{E}' = (0, vE' \cos\theta, -vE' \sin\theta) \tag{14.76}$$

and so

$$\begin{aligned} \mathbf{B}' &= k_B \frac{1}{c}(\mathbf{v}' \times \mathbf{E}') \\ &= k_B(0, \beta E' \cos\theta, -\beta E' \sin\theta). \end{aligned} \tag{14.77}$$

Varying the angle θ shows that this is directed in a circle about the moving line of charge, as shown in fig. 14.6. The magnitude of $\mathbf{B}'$ is

$$B' = k_B \beta E' = \left(\frac{k_B k_E}{c}\right) \frac{\lambda' v}{2\pi r} = \left(\frac{k_B k_E}{c}\right) \frac{i'}{2\pi r}. \tag{14.78}$$

The proportionality constant depends on the system of units:

$$\left(\frac{k_B k_E}{c}\right) = \begin{cases} \frac{4\pi}{c} & \text{Gaussian (cgs) units} \\ \frac{1}{c^2 \epsilon_0} \equiv \mu_0 & \text{SI units} \end{cases} \tag{14.79}$$

The quantity μ_0 in SI units is known as the magnetic permeability (of vacuum).

14.8.1 Current bearing wires

An ordinary electric current in a metallic wire consists of the movement of negatively charged electrons. The electrons move with a "drift velocity" of order of a few mm/s. Also in the wire are stationary positive charges, residing on the atoms in the metal from which the electrons came.

In the rest frame of the wire, the wire is electrically neutral. The separation between the electrons is not Lorentz contracted because there are only a finite number of electrons in the wire and charge is conserved. Thus the electrons remain spread out over the entire wire.

Let the linear charge density of positive charges be λ_+ and that of the electrons be λ_-. Because the wire is electrically neutral we have

$$\lambda_+ = -\lambda_- \equiv \lambda_0 \tag{14.80}$$

so that the net charge density is

$$\lambda = \lambda_+ + \lambda_- = 0. \tag{14.81}$$

Thus the electric field about the wire is zero and a test charge will experience no electric force.

There will however be a magnetic field surrounding the wire due to the moving electrons. If the drift velocity of the electrons is v then by Eqs. 14.78 and 14.61 the magnetic field is of magnitude

$$B = \left(\frac{k_B k_E}{c}\right) \frac{\lambda_0 v}{2\pi r}. \tag{14.82}$$

Thus a moving test charge q will experience a magnetic force. If the test charge is moving parallel to the wire with the same speed, v, as the electrons then the magnitude of this force is

$$F_m = q\frac{\beta}{k_B}B = q\beta^2 k_E \frac{\lambda_0}{2\pi r}. \tag{14.83}$$

One observes that this is just β^2 times what the electric force on the test charge would be if there were only positive charges in the wire, i.e.

$$F_e = qE = qk_E\frac{\lambda_0}{2\pi r}. \tag{14.84}$$

This is no accident.

Consider the forces on the test charge in its own rest frame (and that of the electrons). In this frame there is no magnetic force. However there is an electric force because in this frame the wire is not electrically neutral! To see this, consider the charge densities of the positive charges and the electrons. We have

$$\lambda'_+ = \frac{Q_+}{l'_+} = \gamma\frac{Q_+}{l_+} = \gamma\lambda_+ = \gamma\lambda_0 \tag{14.85}$$

because of Lorentz contraction of the moving positive charges. However, in this frame the electrons are stationary and their separation is a proper length. Thus

$$\lambda'_- = \frac{Q_-}{l'_-} = \frac{1}{\gamma}\frac{Q_-}{l_-} = \frac{\lambda_-}{\gamma} = -\frac{\lambda_0}{\gamma}. \tag{14.86}$$

This is depicted schematically in fig. 14.7. The net charge density is therefore

$$\lambda' = \lambda'_+ + \lambda'_- = (\gamma - \frac{1}{\gamma})\lambda_0 = \beta^2\gamma\lambda_0. \tag{14.87}$$

Hence, in the rest frame of the test charge there is an electric field

$$E' = k_E\frac{\lambda'}{2\pi r} = \beta^2\gamma k_E\frac{\lambda_0}{2\pi r}. \tag{14.88}$$

The electric force on the test charge is

$$F'_e = qE' = q\beta^2\gamma k_E\frac{\lambda_0}{2\pi r}, \tag{14.89}$$

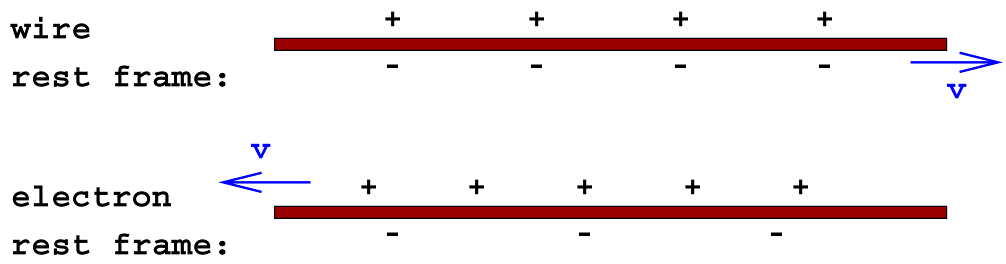

Figure 14.7: A current bearing wire is electrically neutral in its rest frame but appears to carry a net charge in other frames, such as a frame where the conduction electrons are at rest.

pointing radially away from the wire. If we transform this force back to the rest frame of the wire, using the inverses of Eqs. 13.36 or 13.37 with $u'_x = 0$, we get

$$F = \frac{F'_e}{\gamma} = q\beta^2 k_E \frac{\lambda_0}{2\pi r} \tag{14.90}$$

and this is precisely the magnetic force found earlier.

So the magnetic force experienced by the test charge is correctly understood as arising from an electric force in another frame due to a line of net charge density $\gamma\beta^2\lambda_0$, which would correspond to a effective charge density $\beta^2\lambda_0$ in the rest frame of the wire. The relative strength of the magnetic and electric forces is thus of the order of β^2 where β is determined by the drift velocity of the electrons. As remarked above, this is of the order of just a few mm/s and so

$$\beta \approx 10^{-11} \tag{14.91}$$

for typical current bearing wires. The magnetic force they exert is therefore just 10^{-22} times that of an electric force due to just the positive (or negative) charges. That the magnetic force is observed at all (and is appreciable) is due to the enormous amount of charge carried by a number of electrons of the order of Avogadro's number. That this magnetic force is not completely swamped by a still larger electric force is due to the exact cancellation of the positive and negative charges in the wire.

14.8.2 Forces between currents in wires

Suppose two current bearing wires are parallel to each other.

Currents in same direction

Consider the forces on the positive charges in the rest frame of the wires. Since the positive charges are at rest they experience no magnetic force and since each wire is electrically neutral, the positive charges in one will experience no net electric force due to the charges in the other.

Now consider the negative charges. Again, they experience no electric force. They do though experience a magnetic force because they are moving in the magnetic field of the moving negative charges in the other wire. Rather than calculate this magnetic force we note that it arises due to an electric force in a frame in which the electrons are at rest. In that frame each wire appears to possess a net positive charge, as discussed above. (See fig. 14.8.) So the negative charges in one wire experience an attractive electric force towards the other wire.

One concludes therefore that the wires will attract each other and this attraction must take place in any frame, regardless of the combination of electric and magnetic effects to which we attribute it.

Currents in opposite direction

As above, the positive charges will experience no force of any kind in the rest frame of the wires. The negative charges though will experience a purely magnetic force that can be understood as arising from an electric force in another frame.

Consider a frame in which the negative charges in one wire are at rest. They will therefore experience a purely electric force. In the other wire, both positive and negative charges are moving (and in the same direction). The negative ones will be moving faster than the positive ones since they were already moving in that direction in the original frame. Thus the contraction factor for the electrons is greater than that for the positive charges and this wire will present a net negative charge density. (See fig. 14.9.) Accordingly the electrons at rest in the other wire will be repelled. This repulsive electric force corresponds to a repulsive magnetic force in the original frame and the wires will repel each other.

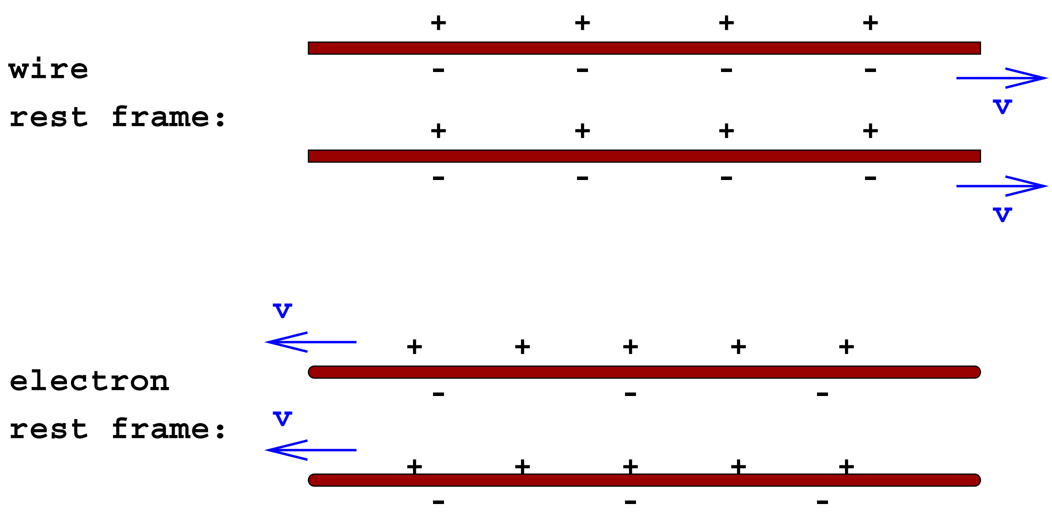

Figure 14.8: The magnetic force between conduction electrons in the rest frame of two wires may be interpreted as an attractive electric force on those electrons in their rest frame.

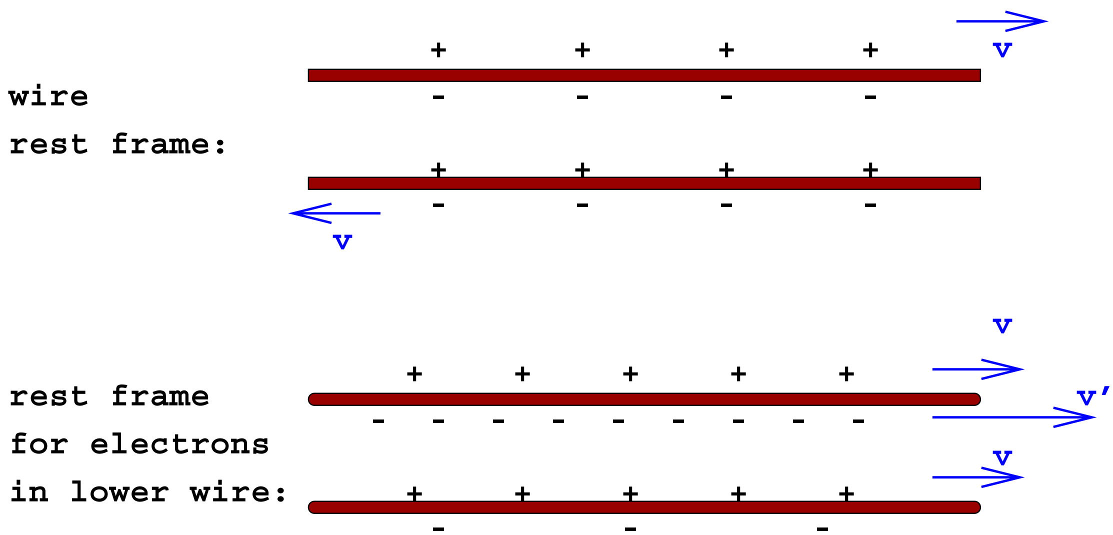

Figure 14.9: The magnetic force between conduction electrons in the rest frame of two wires carrying current in opposite directions may be interpreted as a repulsive electric force on those electrons in their rest frame.

14.9 Maxwell's equations

The theory of electromagnetism is traditionally summarized in terms of the Lorentz force and four laws, known as Maxwell's equations. We show in this section how the four Maxwell equations can be obtained from the basic phenomenological laws of electrostatics together with relativistic transformations.

14.9.1 Gauss's law for electric fields

The first of Maxwell's equations has already been encountered. It is Gauss's law for electric fields, Eq. 14.5:

$$\oint \mathbf{E} \cdot d\mathbf{A} = k_E Q_{\text{enc}}.$$

As we have already remarked, this result is true in any inertial frame.

Sometimes it is useful to rewrite this law in another way. There is a mathematical theorem, known as the divergence theorem [26], which states that for any vector field, $\mathbf{E}$,

$$\oint_S \mathbf{E} \cdot d\mathbf{A} = \int_V \nabla \cdot \mathbf{E}\, d\tau \tag{14.92}$$

where S is any closed surface and $d\tau$ is an element of the volume, V, enclosed by that surface. The quantity

$$\nabla \cdot \mathbf{E} \equiv \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right) \cdot (E_x, E_y, E_z) = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \tag{14.93}$$

is known as the divergence of $\mathbf{E}$ and the theorem thus states that the integral of $\mathbf{E}$ over the surface S is the same as the integral of the divergence of $\mathbf{E}$ over the volume, V, enclosed by S. One now notes that the definition of charge density, ρ, see Eq. 14.64, means that

$$Q_{\text{enc}} = \int_V \rho\, d\tau. \tag{14.94}$$

Thus Gauss's law, Eq. 14.5, may be equivalently written as

$$\int_V \nabla \cdot \mathbf{E}\, d\tau = k_E \int_V \rho\, d\tau. \tag{14.95}$$

Since V is a completely arbitrary volume though, this requires that

$$\nabla \cdot \mathbf{E} = k_E \rho \tag{14.96}$$

at every point in space. This result is known as Gauss's law (for electric fields) in differential form, as opposed to the earlier Gauss's law in integral form.

14.9.2 Gauss's law for magnetic fields

The second of Maxwell's equations involves an analogous closed surface integral for the magnetic field, i.e. $\oint \mathbf{B} \cdot d\mathbf{A}$. In order to determine the value of this integral we first note that the magnetic field can be broken up into the separate contributions of individual moving charges, which are then added vectorially. We can therefore consider just those charges that are moving with a common velocity, $\mathbf{v}$, and then add their contribution to those of charges with other velocities. Also, the volume inside the (arbitrary) closed surface of integration can be broken up into a series of many small cubic volumes. Since $\mathbf{B}$ must be in a fixed direction within a small region, the value of the magnetic flux

$$d\Phi_B = \mathbf{B} \cdot d\mathbf{A} \tag{14.97}$$

will cancel on adjacent surfaces of the cubes since $d\mathbf{A}$ is opposite for adjacent sides of neighboring cubes. Thus, upon summing the closed surface integrals over individual cubes, only the integrals over faces exterior to the complete volume (where there are no other cubes adjacent) will survive the summation. By making the cubes as small as necessary it is clear that the sum of integrals over the cubes can be made to approximate the closed surface integral in question to any desired accuracy. We thus focus on a single integral over an arbitrary cube.

We next note that the magnetic field produced by moving charges is perpendicular to $\mathbf{v}$. Without loss of generality, we choose the x-axis to be in the direction of $\mathbf{v}$ and align all cubes with the coordinate axes. In that case $\mathbf{B} \cdot d\mathbf{A}$ vanishes over the two surfaces that are perpendicular to the x-axis and we need only consider the remaining integrals over the four cube faces that are parallel to the direction of $\mathbf{v}$. See fig. 14.10.

The remaining contributions to the integral may be evaluated by considering the frame in which the charges are at rest. In that frame there is

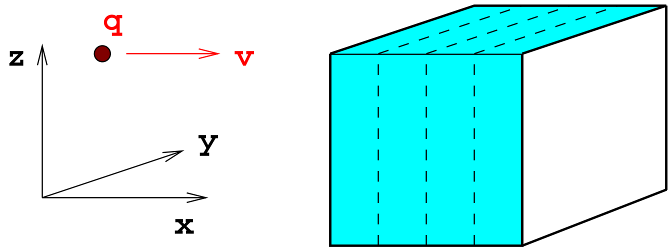

Figure 14.10: The magnetic flux is non-zero only over the shaded surfaces lying parallel to the direction of charge motion.

no magnetic field and the transformation, Eq. 14.59, for the magnetic field reduces to

$$\mathbf{B}'_{\perp} = -\gamma \frac{k_B}{c} \mathbf{v} \times \mathbf{E} \tag{14.98}$$

where $\mathbf{E}$ is the electric field in the charges' rest frame and $\mathbf{v}$ is the relative velocity of the frames. Thus

$$\begin{aligned} \int_{\substack{\text{surfaces} \\ \text{parallel to } \mathbf{v}}} \mathbf{B}' \cdot d\mathbf{A}' &= -\frac{\gamma k_B}{c} \int_{\substack{\text{surfaces} \\ \text{parallel to } \mathbf{v}}} (\mathbf{v} \times \mathbf{E}) \cdot d\mathbf{A}' \\ &= -\frac{\gamma^2 k_B}{c} \int_{\substack{\text{surfaces} \\ \text{parallel to } \mathbf{v}}} (\mathbf{v} \times \mathbf{E}) \cdot d\mathbf{A}. \end{aligned} \tag{14.99}$$

where $d\mathbf{A}$ is a moving surface in the charges' rest frame (viewed instantaneously in that frame). Bearing in mind that $d\mathbf{A}$ points outwards from the cube surfaces we have, with reference to fig. 14.10

$$d\mathbf{A} = \begin{cases} -dxdz\, \mathbf{e}_y & \text{on near surface} \\ +dxdy\, \mathbf{e}_z & \text{on top surface} \\ +dxdz\, \mathbf{e}_y & \text{on far surface} \\ -dxdy\, \mathbf{e}_z & \text{on bottom surface.} \end{cases} \tag{14.100}$$

We also note that $\mathbf{v} = (v, 0, 0)$. Thus

$$\mathbf{v} \times \mathbf{E} \cdot d\mathbf{A} = \begin{cases} +vE_z^{(\text{near})}\, dxdz & \text{on near surface} \\ +vE_y^{(\text{top})}\, dxdy & \text{on top surface} \\ -vE_z^{(\text{far})}\, dxdz & \text{on far surface} \\ -vE_y^{(\text{bottom})}\, dxdy & \text{on bottom surface.} \end{cases} \tag{14.101}$$

Hence, when we integrate we get (for a cube of dimensions Δx, Δy, Δz):

$$\begin{aligned} \int_{\substack{\text{surfaces}\\ \text{parallel to } \mathbf{v}}} & \mathbf{v} \times \mathbf{E} \cdot d\mathbf{A} \\ = & \int_x^{x+\Delta x} dx \underset{(\text{near})}{\int_z^{z+\Delta z}} dz\, vE_z^{(\text{near})} + \int_x^{x+\Delta x} dx \underset{(\text{top})}{\int_y^{y+\Delta y}} dy\, vE_y^{(\text{top})} \\ & - \int_x^{x+\Delta x} dx \underset{(\text{far})}{\int_z^{z+\Delta z}} dz\, vE_z^{(\text{far})} - \int_x^{x+\Delta x} dx \underset{(\text{bottom})}{\int_y^{y+\Delta y}} dy\, vE_y^{(\text{bottom})} \\ = & \; v \int_x^{x+\Delta x} dx \left[\underset{(\text{near})}{\int_z^{z+\Delta z}} dz\, E_z^{(\text{near})} + \underset{(\text{top})}{\int_y^{y+\Delta y}} dy\, E_y^{(\text{top})} \right. \\ & \left. + \underset{(\text{far})}{\int_{z+\Delta z}^{z}} dz\, E_z^{(\text{far})} + \underset{(\text{bottom})}{\int_{y+\Delta y}^{y}} dy\, E_y^{(\text{bottom})} \right] \\ = & \; v \int_x^{x+\Delta x} dx \oint_{\substack{\text{loop at}\\ \text{constant}\\ x}} \mathbf{E} \cdot d\mathbf{l}\,. \end{aligned} \tag{14.102}$$

The closed loop integral here is around loops oriented perpendicularly to the x-axis, as indicated by the dashed lines in fig. 14.10. Since the electric field is that in the charge rest frame, we can apply Eq. 14.4 for conservative fields, showing that this loop integral is zero. The entire integral thus vanishes, for all cubes, independently of the velocity $\mathbf{v}$ of the charges that give rise to $\mathbf{B}$. Since Eq. 14.98 can be used to relate $\mathbf{B}$ in any inertial frame to an electric field in the charges' rest frame we have that

$$\oint \mathbf{B} \cdot d\mathbf{A} = 0 \tag{14.103}$$

for all inertial frames. This is Gauss's law for magnetism.

Just as with Gauss's law for electric fields, this law can be rewritten in differential form.

Exercise 14.4 *Show that*

$$\nabla \cdot \mathbf{B} = 0 \tag{14.104}$$

everywhere.

14.9.3 Faraday's induction law

The next of Maxwell's equations is a generalization of a result obtained earlier. Recall that for a static electric field (i.e. one arising from stationary source charges) that

$$\oint \mathbf{E} \cdot d\mathbf{x} = 0.$$

This result arose from the fact that the electric force exerted by stationary charges is conservative. We also saw though that the result was not true if the source charges were moving (recall Exercise 14.3). The correct generalization is obtained from the relativistic transformations, Eq. 14.59.

We begin by breaking up the electric field into contributions, $\mathbf{E}^{(v)}$, from those charges moving with a given velocity, $\mathbf{v}$. For each of the values of $\mathbf{v}$ present amongst the charges we apply a (passive) boost transformation to the frame in which those charges (all moving with the same $\mathbf{v}$ in the original frame) are at rest. In this frame the electric field, $\mathbf{E}'^{(0)}$, (arising from these charges) satisfies

$$\oint \mathbf{E}'^{(0)} \cdot d\mathbf{x}' \equiv \oint \left(E'^{(0)}_{\parallel} dx_{\parallel}{}' + \mathbf{E}'^{(0)}_{\perp} \cdot d\mathbf{x}_{\perp}{}' \right) = 0. \tag{14.105}$$

This is true for any closed loop integral. The generality of this result is crucial in what follows. Since $\mathbf{E}'^{(0)}$ is static in this frame it makes no difference if we evaluate the field around the loop at a fixed instant or at different times.

It also doesn't matter if the loop in this particular integral is moving, as the electric field is independent of the velocity of a test charge (residing on the loop). Furthermore, we may even evaluate the field at different times around a moving loop, provided only that those times are such that the loop is closed. The (static) electric field in this transformed frame is given by Eq. 14.59 for a passive transformation by $+\mathbf{v}$:

$$\begin{aligned} E'^{(0)}_{\parallel} &= E^{(v)}_{\parallel} \\ \mathbf{E}'^{(0)}_{\perp} &= \gamma\left(\mathbf{E}^{(v)}_{\perp} + \frac{1}{k_B c}(\mathbf{v}\times\mathbf{B}^{(v)})_{\perp}\right). \end{aligned}$$

This transformation takes a field configuration at some point and maps it to the field configuration at the transformed point in the new frame. It is important to appreciate that in the (original) frame with moving source charges, the fields are not static and therefore it only makes sense to perform a closed loop integral at a given instant of time. Therefore $dt = 0$ as we move around the loop and the Lorentz transformations for the distance element, Eq. 10.7, reduce to

$$dx'_{\parallel} = \gamma dx_{\parallel},\ d\mathbf{x}'_{\perp} = d\mathbf{x}_{\perp}. \tag{14.106}$$

The image of a stationary loop at a fixed instant in the frame with moving sources is a moving loop at different times in the frame with stationary sources. However, as noted above, this does not matter and we can combine the transformations to obtain:

$$\begin{aligned} 0 = \oint \mathbf{E}'^{(0)}\cdot d\mathbf{x}' &= \oint\left[E^{(v)}_{\parallel}\gamma dx_{\parallel} + \gamma\left(\mathbf{E}^{(v)}_{\perp} + \frac{1}{k_B c}(\mathbf{v}\times\mathbf{B}^{(v)})_{\perp}\right)\cdot d\mathbf{x}_{\perp}\right] \\ &= \gamma\oint\left[\left(E^{(v)}_{\parallel}dx_{\parallel} + \mathbf{E}^{(v)}_{\perp}\cdot d\mathbf{x}_{\perp}\right) + \frac{1}{k_B c}(\mathbf{v}\times\mathbf{B}^{(v)})_{\perp}\cdot d\mathbf{x}_{\perp}\right] \\ &= \gamma\left[\oint \mathbf{E}^{(v)}\cdot d\mathbf{x} + \frac{1}{k_B c}\oint(\mathbf{v}\times\mathbf{B}^{(v)})_{\perp}\cdot d\mathbf{x}_{\perp}\right] \end{aligned} \tag{14.107}$$

Therefore,

$$\oint \mathbf{E}^{(v)}\cdot d\mathbf{x} = -\frac{1}{k_B c}\oint(\mathbf{v}\times\mathbf{B}^{(v)})_{\perp}\cdot d\mathbf{x}_{\perp}. \tag{14.108}$$

Since $(\mathbf{v}\times\mathbf{B}^{(v)})$ is a vector perpendicular to $\mathbf{v}$, we must have

$$(\mathbf{v}\times\mathbf{B}^{(v)})\cdot d\mathbf{x}_{\parallel} = 0 \tag{14.109}$$

and thus

$$\oint \mathbf{E}^{(v)} \cdot d\mathbf{x} = -\frac{1}{k_B c} \oint (\mathbf{v} \times \mathbf{B}^{(v)}) \cdot d\mathbf{x}. \tag{14.110}$$

To proceed we make use of a mathematical result known as Stoke's theorem (or the curl theorem) [26]. This states that for any vector field $\mathbf{V}$,

$$\oint \mathbf{V} \cdot d\mathbf{x} = \int_A (\nabla \times \mathbf{V}) \cdot d\mathbf{A} \tag{14.111}$$

where A is any surface whose boundary is the closed loop appearing in the closed loop integral. The quantity $\nabla \times \mathbf{V}$ is known as the curl of $\mathbf{V}$ and is given by

$$\nabla \times \mathbf{V} = \left(\frac{\partial V_z}{\partial y} - \frac{\partial V_y}{\partial z}, \frac{\partial V_x}{\partial z} - \frac{\partial V_z}{\partial x}, \frac{\partial V_y}{\partial x} - \frac{\partial V_x}{\partial y} \right). \tag{14.112}$$

Applying this result we find that

$$\oint (\mathbf{v} \times \mathbf{B}^{(v)}) \cdot d\mathbf{x} = \int_A \nabla \times (\mathbf{v} \times \mathbf{B}^{(v)}) \cdot d\mathbf{A}. \tag{14.113}$$

The quantity appearing in the surface integral may be evaluated by noting the general mathematical result [26] that

$$\nabla \times (\mathbf{v} \times \mathbf{B}^{(v)}) = (\mathbf{B}^{(v)} \cdot \nabla)\mathbf{v} - \mathbf{B}^{(v)}(\nabla \cdot \mathbf{v}) + \mathbf{v}(\nabla \cdot \mathbf{B}^{(v)}) - (\mathbf{v} \cdot \nabla)\mathbf{B}^{(v)} \tag{14.114}$$

The first two terms vanish because $\mathbf{v}$ is constant and the third vanishes by virtue of Gauss's law for magnetic fields, Eq. 14.104. The remaining term is

$$-(\mathbf{v} \cdot \nabla)\mathbf{B}^{(v)} = -\left(v\frac{\partial}{\partial x} \right) \mathbf{B}^{(v)} \tag{14.115}$$

and gives the rate of change of $\mathbf{B}^{(v)}$ in the direction of $\mathbf{v}$. One recalls that in this frame, $\mathbf{B}^{(v)}$ arises from a moving electric field configuration as in Eq. 14.56 and Fig. 14.5 (but with the charge moving in the positive $\mathbf{x}$ direction in the present case). Thus

$$\mathbf{B}^{(v)}(\mathbf{x} + \Delta\mathbf{x}, t) = \mathbf{B}^{(v)}(\mathbf{x}, t - \Delta t) \tag{14.116}$$

where

$$\Delta t = \frac{\Delta x}{v}.$$

Therefore

$$\begin{aligned} v\frac{\partial \mathbf{B}^{(v)}}{\partial x} &\equiv v \lim_{\Delta x \to 0} \frac{\mathbf{B}^{(v)}(x+\Delta x, y, z, t) - \mathbf{B}^{(v)}(x, y, z, t)}{\Delta x} \\ &= \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} \frac{\mathbf{B}^{(v)}(x, y, z, t-\Delta t) - \mathbf{B}^{(v)}(x, y, z, t)}{\Delta x} \\ &= -\frac{\partial \mathbf{B}^{(v)}}{\partial t}. \end{aligned} \tag{14.117}$$

(Note that since the magnetic field in the above derivation is a function of four variables, x, y, z and t, naive application of the chain rule for differentiation is wrong.) Since $d\mathbf{A}$ is itself constant we end up with

$$\begin{aligned} \oint (\mathbf{v} \times \mathbf{B}^{(v)}) \cdot d\mathbf{x} &= -\int_A \big((\mathbf{v} \cdot \nabla)\mathbf{B}^{(v)}\big) \cdot d\mathbf{A} \\ &= +\int_A \frac{\partial}{\partial t} \big(\mathbf{B}^{(v)} \cdot d\mathbf{A}\big) \\ &= +\frac{\partial}{\partial t} \int_A \mathbf{B}^{(v)} \cdot d\mathbf{A} \\ &= +\frac{\partial \Phi_B^{(v)}}{\partial t}. \end{aligned} \tag{14.118}$$

Inserting this back in Eq. 14.110 and adding up all of the similar contributions from each $\mathbf{E}^{(v)}$ we finally obtain the general result desired:

$$\oint \mathbf{E} \cdot d\mathbf{x} = -\frac{1}{k_B c} \frac{\partial \Phi_B}{\partial t} \tag{14.119}$$

which is valid at any instant in any inertial frame.

This result is known as Faraday's induction law. It says that the integral of the electric field around a closed loop is proportional to (the negative of) the time rate of change of the magnetic flux through the loop. Since an electric field will cause charges to move, if the loop follows a physical conducting wire then a current will be induced in this wire. This is the basis of the electric motor.

Faraday's law may also be cast in differential form by application of the curl theorem. We have

$$\oint \mathbf{E} \cdot d\mathbf{x} = \int_A \nabla \times \mathbf{E} \cdot d\mathbf{A} = -\frac{1}{k_B c} \int_A \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{A} \tag{14.120}$$

for arbitrary areas A. Hence we must have

$$\nabla \times \mathbf{E} = -\frac{1}{k_B c} \frac{\partial \mathbf{B}}{\partial t} \tag{14.121}$$

everywhere.

14.9.4 Ampère-Maxwell law

The final Maxwell equation is a closed loop integral for the magnetic field, analogous to that for the electric field in Faraday's induction law.

To derive the Ampère-Maxwell law we proceed along the same lines as the derivation of Faraday's induction law in the last section. We first decompose the magnetic field $\mathbf{B}$ into contributions $\mathbf{B}^{(v)}$ from charges moving with velocity $\mathbf{v}$ and then for each such contribution we consider a frame in which those charges are at rest. In this frame the transformed magnetic field is

$$\mathbf{B}'^{(0)} = 0. \tag{14.122}$$

Hence

$$\oint \mathbf{B}'^{(0)} \cdot d\mathbf{x}' = 0 \tag{14.123}$$

around any closed loop, including loops which may be moving, and even if the field is evaluated at non-equal times around the loop. Thus, applying the transformation equations 14.106 for a loop at a fixed instant in the frame with moving charges and the transformations, Eq. 14.59, for the magnetic field, i.e.

$$\begin{aligned} B'^{(0)}_{\parallel} &= B^{(v)}_{\parallel} \\ \mathbf{B}'^{(0)}_{\perp} &= \gamma \left(\mathbf{B}^{(v)}_{\perp} - \frac{k_B}{c} (\mathbf{v} \times \mathbf{E}^{(v)})_{\perp} \right), \end{aligned}$$

we obtain

$$0 = \oint \mathbf{B}'^{(v)} \cdot d\mathbf{x}' = \gamma \left[\oint \mathbf{B}^{(v)} \cdot d\mathbf{x} - \frac{k_B}{c} \oint (\mathbf{v} \times \mathbf{E}^{(v)})_\perp \cdot d\mathbf{x}_\perp \right]. \quad (14.124)$$

Therefore

$$\begin{aligned} \oint \mathbf{B}^{(v)} \cdot d\mathbf{x} &= +\frac{k_B}{c} \oint (\mathbf{v} \times \mathbf{E}^{(v)})_\perp \cdot d\mathbf{x}_\perp \\ &= +\frac{k_B}{c} \oint (\mathbf{v} \times \mathbf{E}^{(v)}) \cdot d\mathbf{x} \\ &= +\frac{k_B}{c} \int_A \nabla \times (\mathbf{v} \times \mathbf{E}^{(v)}) \cdot d\mathbf{A}. \end{aligned} \quad (14.125)$$

Consider now

$$\nabla \times (\mathbf{v} \times \mathbf{E}^{(v)}) = (\mathbf{E}^{(v)} \cdot \nabla)\mathbf{v} - \mathbf{E}^{(v)}(\nabla \cdot \mathbf{v}) + \mathbf{v}(\nabla \cdot \mathbf{E}^{(v)}) - (\mathbf{v} \cdot \nabla)\mathbf{E}^{(v)}. \quad (14.126)$$

As was the case in the derivation of the Faraday induction law, the first two terms vanish and

$$-(\mathbf{v} \cdot \nabla)\mathbf{E}^{(v)} = +\frac{\partial \mathbf{E}^{(v)}}{\partial t} \quad (14.127)$$

The third term is, using Gauss's law, Eq. 14.96, together with Eq. 14.65

$$\mathbf{v}(\nabla \cdot \mathbf{E}^{(v)}) = \mathbf{v} k_E \rho^{(v)} = k_E \mathbf{j}^{(v)} \quad (14.128)$$

where $\mathbf{j}^{(v)}$ is the current density arising from the charges moving with velocity $\mathbf{v}$. Combining the pieces,

$$\oint \mathbf{B}^{(v)} \cdot d\mathbf{x} = +\frac{k_B}{c} \frac{\partial}{\partial t} \int_A \mathbf{E}^{(v)} \cdot d\mathbf{A} + \frac{k_B k_E}{c} \int_A \mathbf{j}^{(v)} \cdot d\mathbf{A} \quad (14.129)$$

and summing all contributions (from various $\mathbf{v}$) we obtain the general result, valid in any inertial frame, that

$$\oint \mathbf{B} \cdot d\mathbf{x} = \frac{k_B}{c} \frac{\partial \Phi_E}{\partial t} + \frac{k_B k_E}{c} i \quad (14.130)$$

where

$$i = \int_A \mathbf{j} \cdot d\mathbf{A} \quad (14.131)$$

is the current flowing through the loop.

If the electric flux is not changing then the result reduces to one stated by Ampère. The need for the other term was deduced by Maxwell via consideration of what happens when a capacitor is charging, together with some brilliant mathematical deductions. The full result is therefore known as the Maxwell-Ampère law.

This law may also be written in differential form by applying the curl theorem.

Exercise 14.5 *Show that*

$$\nabla \times \mathbf{B} = \frac{k_B}{c}\frac{\partial \mathbf{E}}{\partial t} + \frac{k_B k_E}{c}\mathbf{j}. \tag{14.132}$$

14.10 Electromagnetic waves

While the transformations 14.59 relate electric and magnetic fields in different frames, Faraday's induction law and the Ampère-Maxwell law relate electric and magnetic fields in the same frame. Indeed they form a pair of coupled differential equations. In the general case where we have a region of space with no charges present (and hence also no currents) they reduce to

$$\begin{aligned}\nabla \times \mathbf{E} &= -\frac{1}{k_B c}\frac{\partial \mathbf{B}}{\partial t}\\ \nabla \times \mathbf{B} &= +\frac{k_B}{c}\frac{\partial \mathbf{E}}{\partial t}.\end{aligned} \tag{14.133}$$

Using the mathematical result [26], true for any vector field $\mathbf{V}$, that

$$\nabla \times (\nabla \times \mathbf{V}) = \nabla(\nabla \cdot \mathbf{V}) - \nabla^2 \mathbf{V} \tag{14.134}$$

we can separate these coupled equations by taking the curl of both sides. We find

$$\begin{aligned}\nabla \times (\nabla \times \mathbf{E}) = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} &= -\frac{1}{k_B c}\frac{\partial}{\partial t}\nabla \times \mathbf{B}\\ &= -\frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2}.\end{aligned} \tag{14.135}$$

Since $\nabla \cdot \mathbf{E} = 0$ if no charges are present we have

$$\nabla^2 \mathbf{E} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{E} = 0. \tag{14.136}$$

Similarly,

$$\nabla^2 \mathbf{B} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{B} = 0. \tag{14.137}$$

These equations describe wave oscillations travelling at velocity c.

Exercise 14.6 *Show that*

$$\mathbf{E} = E_0 \cos(x - ct)\mathbf{e}_y \tag{14.138}$$

is one possible solution of the wave equation for $\mathbf{E}$ *and that a maximum of this oscillating solution moves in the positive* x *direction with speed* c*.*

Show further that this oscillating electric field induces an oscillating magnetic field given by

$$\mathbf{B} = +k_B c E_0 \cos(x - ct)\mathbf{e}_z \tag{14.139}$$

which is at right angles to the electric field. Also observe that the oscillations of both electric and magnetic fields are transverse to the direction of propagation.

When Maxwell first deduced these equations, the constants denoted here as $1/k_B c$ and k_B/c were instead expressed in terms of electric and magnetic constants. By inserting the known values of those constants, Maxwell arrived at the unexpected result that oscillating electric and magnetic fields propagated in free space at the same speed as that of light. He therefore surmised that light was an oscillating electromagnetic wave and so unified two previously disparate branches of physics. Experimental confirmation of the existence of electromagnetic waves was achieved by Hertz in 1890 and lead quickly to the development of radio.

The presentation in this chapter shows that it is no coincidence that the propagation speed of electromagnetic waves is that of light. Rather, it follows naturally from relativistic transformations of static electrical phenomena. Indeed, electromagnetic waves themselves should not be regarded as something mysteriously predicted by a complex theory, even though that is bound to be one's first impression. It is to be recalled that magnetic fields only exist in the mind of the beholder — as a convenient representation of the forces that are exerted on test charges if one is unfortunate enough to be in an inertial frame where the source charges are moving. When those forces are represented in this way, the relativistic transformations dictate that changing electric fields induce (changing) magnetic fields and vice versa. In free space those changing fields were seen to take the form of an electromagnetic wave. Furthermore, the wave equations follow from Maxwell's equations in free space and since Maxwell's equations apply in every inertial frame, so do the wave equations for the electromagnetic fields. The electromagnetic wave should therefore be viewed merely as a solution of time varying electromagnetic fields consistent with Lorentz symmetry in free space.

This point of view can be brought into focus by considering the transformation of the derivatives in the wave equation. Recalling the Lorentz transformations, and their inverses, we note that the coordinates in one frame can each be expressed as a function of the coordinates in another. For example $t' = t'(x, y, z, t)$ and so the partial derivative with respect to t' can be written as

$$\frac{\partial}{\partial t'} = \frac{\partial t}{\partial t'}\frac{\partial}{\partial t} + \frac{\partial x}{\partial t'}\frac{\partial}{\partial x} + \frac{\partial y}{\partial t'}\frac{\partial}{\partial y} + \frac{\partial z}{\partial t'}\frac{\partial}{\partial z}. \tag{14.140}$$

Using the (inverse) Lorentz transformations to evaluate the partial derivatives we obtain

$$\frac{1}{c}\frac{\partial}{\partial t'} = \gamma\left(\frac{1}{c}\frac{\partial}{\partial t} + \beta\frac{\partial}{\partial x}\right). \tag{14.141}$$

Similarly,

$$\frac{\partial}{\partial x'} = \gamma\left(\frac{\partial}{\partial x} + \beta\frac{1}{c}\frac{\partial}{\partial t}\right) \tag{14.142}$$

$$\frac{\partial}{\partial y'} = \frac{\partial}{\partial y} \tag{14.143}$$

$$\frac{\partial}{\partial z'} = \frac{\partial}{\partial z}. \tag{14.144}$$

Exercise 14.7 *Use these results to show that*

$$\nabla'^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t'^2} = \nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} \tag{14.145}$$

and hence that the wave equation operator retains its form in every inertial frame.

Review

Do electric charges of the same type repel or attract each other?

Does the value of the charge depend on the reference frame?

What is meant by the electrostatic potential?

What is Gauss's law? Does it depend on the reference frame?

What is the significance of the inverse-square nature of Coulomb's law?

Describe qualitatively the electric field about a point charge, a parallel plate capacitor and a line of charge.

How does the electric field of a moving point charge differ qualitatively from that of a stationary point charge?

What is the essential difference between an electric force and a magnetic force?

What is the Lorentz force law?

How is the magnetic field **B** defined?

How is electric current defined and what does it mean?

Describe qualitatively the magnetic field about a moving line of charge.

What is a typical value for the drift velocity of conduction electrons in a metal wire?

Explain qualitatively how a magnetic field about a current carrying wire is related to an electric field in another frame.

Questions

1. Is there any real benefit in using separate electric and magnetic fields to describe a common phenomenon?

2. Despite the weight of authority behind SI units, cgs units continue to be very widely used in electricity and magnetism. Why do you suppose this might be?

3. It has sometimes been suggested that the theory of electromagnetism would take on a more symmetric and pleasing look if there existed magnetic monopoles, i.e. sources of magnetic fields analogous to the electric charges which are the sources of electric fields. If that were the case then Gauss's law for magnetism would involve the number of monopoles enclosed in the Gaussian volume and the closed surface integral of $\mathbf{B} \cdot d\mathbf{A}$ would no longer be zero. Do you consider that such an addition to the theory would really make it more beautiful? Would electric fields ever be conservative in the presence of magnetic monopoles? Should magnetic monopoles exist?

Problems

1. Show that the electric field inside a hollow sphere carrying a net charge Q on its surface is zero but that the electric field outside is the same as that for a point charge Q located at the center of the sphere.

2. Using the definition of current and the definition of a magnetic force, show that the force on a straight conductor of length L, bearing a current i in a uniform magnetic field is

$$\mathbf{F} = \frac{i}{k_B c} \mathbf{L} \times \mathbf{B} \tag{14.146}$$

   where $\mathbf{L}$ is a vector of magnitude L pointing along the wire in the direction of the current.

3. Use the result of the previous problem to show that the force between two parallel conducting wires of length L, separated by a distance d, is

$$F = \frac{k_E}{c^2} \frac{i_1 i_2 L}{2\pi d}. \tag{14.147}$$

   Determine expressions for the constant k_E/c^2 in both Gaussian and SI units.

4. Using the Lorentz force and the force transformations, establish Eq. 14.59.

5. Consider two parallel copper wires with currents flowing in the same direction. Transform to the frame in which the conduction electrons are at rest. Show explicitly that in this frame, the positive charges in either wire experience both an electric and a magnetic force but that these exactly cancel each other out. Explain, with reference to the rest frame of the wires why this must be so.

6. Consider a square loop of wire in which a current, i, flows. What is the net charge in the wire in its rest frame? Consider now a frame in which the conduction electrons in one side of the loop are at rest. Determine the charge densities in each of the four sides of the loop in terms of the charge densities in the wire's rest frame. Show explicitly that the net charge in the wire is the same as in its rest frame.

7. Show that the electromagnetic 4-force in a frame in which sources are at rest is

$$\mathcal{F}^{\mu} = q\gamma(\mathbf{E} \cdot \beta, \mathbf{E}). \tag{14.148}$$

Explain clearly what β is and whether or not it is directly related to γ.

Chapter 15

Gravity

Objects released near Earth's surface fall towards it. The effect causing this is said to be *gravity*. Newton recognized that the same effect was responsible for controlling the orbital motion of the Moon about Earth, and indeed it governs the motion of all heavenly bodies including the planets, stars and galaxies. Gravity is an interaction occurring between all massive bodies. It can be observed even on a small scale, though the apparent force is very weak.

Newton presented a theory of gravitation in which he described it in terms of a gravitational force between massive objects. The force depended on the mass of the objects and for spherical objects obeyed an inverse-square law:

$$\mathbf{F}_g = -G\frac{Mm}{r^2}\,\mathbf{e}_r \tag{15.1}$$

where M is the mass of the object giving rise to the gravitational force $\mathbf{F}_g$ acting on the mass m, a distance r away, and G is the *universal gravitational constant.* The unit vector $\mathbf{e}_r$ points radially outward from M and the minus sign implies attraction.

Exercise 15.1 *Verify, with the aid of a diagram, that Newton's gravitational formula for the force exerted on M by m is consistent with Newton's third law (Eq. 13.18) for a pair of masses in isolation.*

Exercise 15.2 *By integrating radially outward, and choosing the integration constant, show that the gravitational potential energy is described by*

$$U = -G\frac{Mm}{r}. \tag{15.2}$$

One can define, analogously to the electric potential, a gravitational potential, Φ, as the gravitational potential energy per unit test mass. One can also define a gravitational field, g, as the gradient of the gravitational potential (or the force per unit mass).

Exercise 15.3 *Deduce that the gravitational potential about a spherical mass M is*

$$\Phi_{sphere}^{Newton} = -G\frac{M}{r} \tag{15.3}$$

and so

$$\mathbf{F}_g = -m\nabla\Phi. \tag{15.4}$$

Show also that the gravitational field about this spherical mass is

$$g = -\nabla\Phi = -G\frac{M}{r^2}. \tag{15.5}$$

This theory was enormously successful, describing well the orbits of the planets and even successfully predicting the existence and location of the previously unknown planet Neptune, to account for small observed perturbations in the orbit of Uranus. However, very accurate astronomical observations have detected a few shortcomings. In particular, it has been known since 1845 that even after accounting for the perturbations of all the other planets there is a discrepancy of $43''$ of arc per century in the precession of the perihelion of the orbit of the innermost planet, Mercury. Furthermore, Newton's theory has the disturbing feature that the gravitational force implies that distant objects respond instantaneously to changes in position of the masses, contrary to relativistic causality.

In the present chapter we take a brief and rather cursory look at gravity from a different, but very profound, and highly successful, point of view forwarded by Einstein. Einstein's theory reduces to Newton's in the limit of small velocities and weak fields. The full theory is non-linear and very complex so we shall only look at the first order corrections to Newton's theory. This is adequate for most purposes.

15.1 The equality of gravitational and inertial mass

If one drops two objects, of different composition, from the same height, then they will reach the ground together. Some complications can arise from air resistance if the objects are very light, such as with a feather or a leaf. This may have confused Aristotle who wrote erroneously that heavier objects fall faster. In fact it has been known for a long time that, everything else being equal, objects all fall at the same rate. Ioannes Philiponos, writing in the fifth or sixth century A.D. correctly recorded that even when the difference in weights of the objects was very large, the difference in time of fall was very small.

By the 16th century several people had noted the equality of the fall times and even conducted experiments verifying it. One was Simon Stevin who published his results in 1586, in the time of Galileo. Galileo did further work on the problem and it is part of science folklore that he dropped objects from the leaning tower of Pisa to demonstrate the equality of fall rates. However, he made more accurate measurements by rolling balls down an inclined plane and also by comparing the periods of pendulums with bobs of different material. The latter he found to be the most accurate. In 1638 he wrote that when he used a lead ball and a cork ball to form two pendulums of the same length and started them swinging in unison, they kept perfect synchronization: "neither in a hundred oscillations, nor in a thousand, will the [heavy lead body] anticipate the [light cork body] by even an instant, so perfectly do they keep in step." His claim would actually seem to be an exaggeration because after a thousand oscillations the differing effects of air resistance begin to accumulate.

Newton [11] conducted further experiments with pendulums and was careful to render the effects of air resistance the same for different materials by placing them in two identical wooden boxes. These he suspended from strings 11 feet long and filled with equal weights of different materials such as gold, silver, lead, wood, sand, glass, water, salt and wheat. Swinging the two pendulums together he found they kept in unison for a long time. This is actually quite remarkable. The classical period of a pendulum is derived by equating the component of the gravitational force in the direction of swing with the inertial force according to the low velocity limit of Newton's Second Law. See fig. 15.1. There is no a priori reason why the mass involved

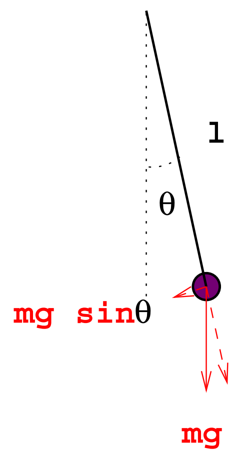

Figure 15.1: A pendulum in a gravitational field g. The component of the gravitational force perpendicular to the string is unbalanced and causes the pendulum to swing.

in Newton's gravitational force should be the same as the mass involved in Newton's second law. Newton envisaged the gravitational force as depending on the amount of matter giving rise to it and we could simply call this the gravitational mass, m_G. The mass in Newton's second law on the other hand, has its origins in the symmetry of spacetime and dictates how a body responds when a force acts on it. We can call it the inertial mass, m_I. These two masses could in principle be different. The electric charge in Coulomb's law, for example, is certainly not equal to the mass of the object carrying it. Thus one has (for small velocities)

$$(m_G g)\sin\theta = m_I a \tag{15.6}$$

and one finds that (for small oscillations) the pendulum exhibits simple harmonic motion with period

$$T = 2\pi\sqrt{\frac{m_I}{m_G}\frac{l}{g}} \tag{15.7}$$

where l is the length of the pendulum and g the gravitational field. Newton's experiments showed that

$$\frac{|m_I - m_G|}{m_I} < 10^{-3}. \tag{15.8}$$

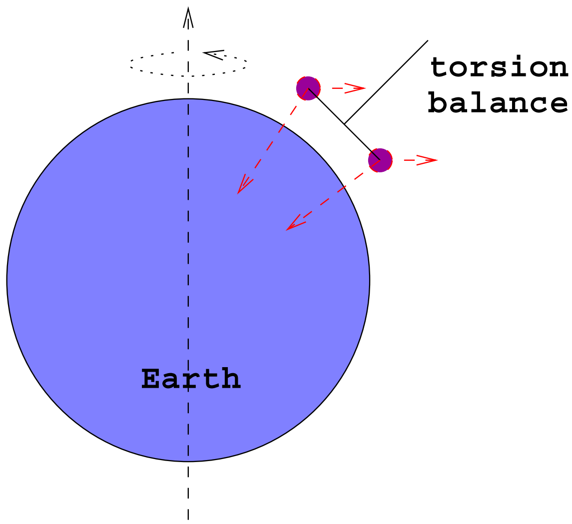

Figure 15.2: The Eötvös experiment: a torsion balance on Earth's surface experiences a gravitational force and an inertial centrifugal force.

Thus, at least in the non-relativistic limit of Newton's theory, the gravitational field is just the acceleration due to gravity.

In 1889 and 1908, the Hungarian Baron Eötvös and collaborators used a torsion balance to improve this limit to less than 3×10^{-9}. The torsion balance uses the principle that gravity will attract a mass towards the center of the Earth. If one places two masses, one on each end of a rod suspended by a thread, then there will, in addition, be a centrifugal force directed away from Earth's spin axis. The strength of this centrifugal force depends on the inertial mass. If the gravitational and inertial masses are not equal then a net torque will exist that will twist the thread. This can be detected by rotating the apparatus by 180° so that the twisting is in the opposite direction.

An interesting refinement of this experiment is due to Dicke [27]. In this experiment one uses the gravitational field of the Sun, and the centrifugal force resulting from Earth's orbital motion, while letting Earth rotate the torsion balance over a period of one day. The most accurate experiment of this type was performed by Braginsky and Panov [28] in 1971. They obtained a limit of less than 9×10^{-13}.

It is to be noted that although all material is composed of protons and neutrons (and electrons), the amount of stored electric and nuclear energy varies from material to material. These experiments show that all types of energy fall in the same way in a gravitational field. Laser ranging measurements to the Moon show that even gravitational energy itself falls at the same rate. If this were not the case, there would be a slight distortion of Moon's orbit about Earth caused by their slightly different accelerations towards the Sun. This is known as the Nordvedt effect and it is not found [29].

Why should bodies of different material composition all fall at the same rate in a gravitational field? Einstein's proposal was that the rate of fall really had nothing directly to do with the mass.

15.1.1 Equivalence principle

Consider oneself inside an enclosed laboratory on the surface of the Earth. If one tosses a ball, it will fall to ground with constant acceleration, following a parabolic trajectory. According to the experiments just discussed, it doesn't matter what kind of ball it is (baseball, tennis ball, cricket ball or lead shot); provided that the initial velocity of the toss is the same, the trajectory and rate of fall will always be identical. But this is precisely what one would observe if one were in an enclosed laboratory that was being accelerated, in the absence of gravitational effects, in the direction of the ceiling. See fig. 15.3. The tossed balls are not connected to the walls of the laboratory and would therefore act as if in an inertial frame, remaining where they were, or retaining the velocity imparted when released, as required by Newton's First Law. It is only because of the non-uniform motion of this laboratory that they would appear to "fall" towards the floor. Furthermore, all types of ball would "fall" at the same rate since the apparent motion is due entirely to the accelerating frame. In fact, even a ray of light would appear bent. If the acceleration of this laboratory was g, the same as the acceleration "due to gravity" at Earth's surface, then there is no way that an observer inside such a laboratory could determine whether he was on the surface of the Earth, witnessing the effects of gravity, or whether he was in remote outer space, far removed from all measurable gravitational effects, and simply being accelerated.

This equality of the two situations is known as Einstein's Equivalence Principle. It provides an elegant explanation of the equality of gravitational and inertial mass. Of course, one must be wary of a laboratory that is large in comparison with any non-uniformity of the gravitational field. In that case tidal effects would be observable and a distinction could be made between

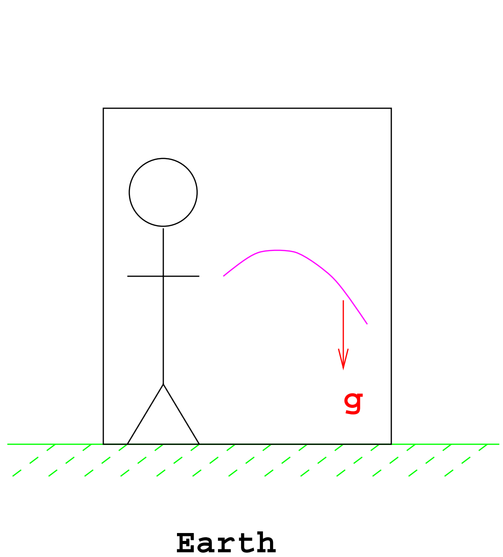

Figure 15.3: The equivalence principle: an observer inside a (small) enclosed laboratory cannot tell the difference between being at rest on Earth's surface or being accelerated in outer space.

the two cases. Therefore, we only apply the Equivalence Principle to suitably small laboratories.

It is important to understand that if we are inside an accelerating laboratory then we are not in an inertial frame. Objects released at rest do not remain at rest with respect to the surrounding laboratory. (This is an experimental fact, debated since the time of Newton, and the philosophical implications of which are still unclear but have been attributed by Mach to the combined gravitational attraction of all the matter in the universe.) Consequently, none of the physics of inertial frames is applicable, including such basic principles as time and length measurement via a system of synchronized clocks. Not only are the methods for clock synchronization not applicable in a non-inertial frame but a system of synchronized clocks does not even exist. However, suppose we are an observer sitting on an object that is released. Then every other object (in the immediate vicinity) released at the same time is going to fall at precisely the same rate with respect to the laboratory and therefore will remain at rest with respect to us. Hence, we would be in a local inertial frame. It follows that all objects in free fall in a gravitational field can be considered as in local inertial frames. In such freely falling (local) inertial frames, we can legitimately apply all of the physics we have learned about (local) inertial frames.

15.2 Gravitational time dilation

Consider a pair of standard clocks in a gravitational field. For definiteness, let clock 1 be on the surface of the Earth and clock 2 be at the top of a mountain, or tall building. Let us now consider the rate at which these clocks will run, assuming that they are identical in every way, apart from their location.

In order to analyze this situation, let us observe the clocks from a laboratory in free fall, i.e. a local inertial frame. From the point of view of this inertial frame, the clocks on Earth are accelerating past. We noted in chapter 6 that an accelerating clock kept the same time as one at the same location, moving with uniform velocity at the same instantaneous speed as the accelerating clock. At least this was true for centripetal accelerations and we adopt the clock hypothesis — as we did when discussing the twin paradox — that this is generally true for all accelerations. Thus, time, as measured by both clock 1 and clock 2, will appear dilated in comparison with the (proper) time τ recorded by a clock in the local inertial frame. (It is still legitimate for a non-inertial observer in the gravitational field to consider his wristwatch as also measuring proper time. But, just as in the twin paradox, only the inertial observer in free fall can readily determine the rate at which the other (non-inertial) observer's clocks run.)

It is important at this point to appreciate the limitations of a *local* inertial frame. In earlier chapters, we have used synchronized clocks at different locations in the same inertial frame. However, the limited extent of a local inertial frame does not permit us to take simultaneous measurements at the different locations of clocks at widely different elevations in a gravitational field. We must wait until we pass each in turn in our local inertial frame. As the local inertial frame falls, it picks up speed. Thus, its speed, v_1, as it passes clock 1 is greater than its speed, v_2, when it passed clock 2. The time dilation factors, γ_1 and γ_2, are therefore different for the two clocks. Specifically,

$$\begin{aligned} t_1 &= \gamma_1 \tau && (15.9) \\ t_2 &= \gamma_2 \tau. && (15.10) \end{aligned}$$

Hence, clock 1 on Earth is seen to run slower than clock 2 at higher elevation.

Similar arguments show this result to be generally true. A clock deep in a gravitational field runs slow compared to one that is far from the massive object about which the field exists. We can obtain an estimate of the size of this effect from the work-energy theorem. From the definition of gravitational potential (see in particular Eqs. 13.60 and 15.4) we see that the infinitesimal amount of work done by the gravitational field in accelerating a particle is

$$dW = -m\nabla\Phi \cdot d\mathbf{x} = -md\Phi \tag{15.11}$$

and thus, using the work-energy theorem,

$$-m(\Phi_2 - \Phi_1) = K_2 - K_1 = mc^2\left(\gamma_2 - \gamma_1\right) , \tag{15.12}$$

i.e.

$$\Delta\gamma = -\frac{\Delta\Phi}{c^2}. \tag{15.13}$$

Since the gravitational potential may be taken as zero a long way from the massive body, and at that point the initial kinetic energy may be selected as zero (corresponding to $\gamma = 1$), we see that a body falling inward (whether radially or otherwise) will acquire the kinetic energy

$$K = -m\Phi^{(0)} \tag{15.14}$$

and thus we readily find that the time dilation factor is

$$\gamma = 1 - \frac{\Phi^{(0)}}{c^2} > 1. \tag{15.15}$$

Exercise 15.4 *Show this. Also show that if the potential is defined as being* $-c^2$ *where the test mass has zero velocity then*

$$\gamma = -\frac{\Phi^{(1)}}{c^2}.$$

(This is actually a quite common choice for the potential since in Einstein's theory, rest energy can also be converted to gravitational potential energy.)

While these equations for the time dilation factor are exact as they stand, their practical implementation begs the question: what is Φ? The derivation uses a potential defined in an inertial frame but we more commonly use potentials expressed in terms of non-inertial coordinate systems. The Newtonian potential, which is only an approximation in Einstein's theory, is such an example. In addition to the issue of what constitutes a valid non-inertial coordinate system, this also raises the general issue of what is meant by a potential, and even energy (or mass too for that matter) in a non-inertial system. Einstein's theory can be viewed as an attempt to define all of these concepts in a self-consistent manner. For the present purposes, since we expect Einstein's theory to reduce to Newton's for weak gravitational fields, we shall simply insert the Newtonian potential so that for weak fields about a spherical mass, M,

$$\gamma_{\text{at r}} \approx 1 + G\frac{M}{rc^2}\,. \tag{15.16}$$

Direct experimental verification of this effect was obtained in the experiment of Hafele and Keating [10] in which cesium-beam atomic clocks were flown around the world. After correction for the kinematic time dilation, the gravitational time dilation was confirmed to within about 10%. Further confirmation was obtained in a slightly more accurate experiment by Alley

[30]. In this experiment an atomic clock was flown in a holding pattern and continuously tracked by laser to determine height and speed.

The most precise test to date has been that of Vessot et al. [31]. In this experiment a highly accurate hydrogen-maser clock was launched on a scout rocket to an elevation of about 10,000 km and the rate was compared via radio signals with a similar clock on the ground. This experiment confirmed the gravitational time dilation to within 2 parts in 10^4.

Global positioning system

The global positioning system makes use of a constellation of 24 satellites orbiting Earth at a height of about 20,000 km. Each satellite contains 4 atomic clocks. This number of satellites is required so that a sufficient number of them will always be visible from anywhere on Earth. By timing radio signals from three different satellites it is possible to use triangulation to pinpoint one's position anywhere. In fact one can get by with only two if one is on Earth's surface and uses this fact to fix one's height. An additional satellite, or more, in addition to several radio frequencies, is used to correct for the propagation delays of the signal through Earth's atmosphere. In order to pinpoint position accurately, very precise timing is required. Ingenious methods are used to reduce the accuracy of clock needed in a GPS receiver but the satellite clocks must be very precise and exactly synchronized with each other. Maintaining this synchronization, by comparison with reference clocks on the ground, requires taking account of relativistic and gravitational corrections.

Originally designed with military uses in mind, civilian use of GPS receivers has exploded, as fairly cheap devices, capable of pinpointing position to within a few tens of metres, have become available. They have found a variety of uses, from surveying (and even measuring continental drift), to helping recreational fishermen locate a favorite fishing hole in a lake. In the near future, cheap devices capable of cm accuracy will be readily available enabling many uses, such as collison avoidance in cars.

The intricacies of time are definitely not an obscure academic matter of no practical importance!

15.2.1 Gravitational redshift

An important manifestation of gravitational time dilation is the gravitational redshift. Suppose we were to consider a light source in a gravitational field. The oscillations associated with the frequency of the light source constitute a clock. (Recall that the standard of time is defined in terms of such a clock.) Similarly, a light detector can be used to measure time in terms of the oscillations of the received light.

To an (inertial) observer in free-fall, the light emitted by a source in a gravitational field will still travel at speed c and retain its frequency (at least over distances for which the free-fall frame may be considered as local inertial). However, the frequency is the inverse of the period of oscillation, and since the period appears dilated to the inertial observer, the frequency appears decreased relative to that of an identical source far removed from the gravitational field. If the period of such a source were T_∞ far away from the gravitational potential then its frequency as viewed by the freely falling observer in the potential is thus

$$f = \frac{1}{T} = \frac{1}{\gamma T_\infty} = \frac{f_\infty}{\gamma}\,. \tag{15.17}$$

Considering now what a detector fixed in the gravitational field would observe, we note that the effect on the frequency derives from the effect on the period which we have concluded is dilated because of the clock hypothesis and thus is exactly as just given for a freely falling local inertial frame at the same location. Thus if the detector is deeper in the gravitational field than the source then it's clocks will be running slow relative to those of the source and therefore the light will appear to arrive with an increased frequency, i.e. it will appear blue-shifted. On the otherhand, if the detector is in a weaker part of the field its clocks will be running faster than those of the source and the light will appear to arrive with a reduced frequency, i.e. it will appear redshifted.

The *gravitational redshift* is defined by

$$\begin{aligned} z_g \equiv \frac{\Delta f}{f} &= \frac{f_2 - f_1}{f_1} = \frac{f_2}{f_1} - 1 = \frac{\gamma_1}{\gamma_2} - 1 \\ &= \frac{1 - \frac{\Phi_1^{(0)}}{c^2}}{1 - \frac{\Phi_2^{(0)}}{c^2}} - 1. \qquad (15.18) \end{aligned}$$

$$\tag{15.19}$$

For a weak gravitational field we can approximate

$$\begin{aligned} z_g &\approx \left(1 - \frac{\Phi_1^{(0)}}{c^2}\right)\left(1 + \frac{\Phi_2^{(0)}}{c^2} + \dots\right) - 1 \\ &= 1 + \frac{(\Phi_2^{(0)} - \Phi_1^{(0)})}{c^2} + \cdots - 1 \\ &\approx +\frac{\Delta\Phi^{(0)}}{c^2} = -\nabla\Phi^{(0)}\frac{\Delta h}{c^2} \end{aligned} \tag{15.20}$$

where Δh is the difference in heights in the gravitational field.

From the above, we note that light emitted from the surface of stars must appear red-shifted to an observer far away, such as an astronomer on Earth. The effect is small however, much smaller than typical kinematic doppler shifts arising from stellar motion, and is thus difficult to detect. Careful measurements of Na and K spectral lines from the Sun have confirmed the effect at about the 5% level [32]. Somewhat less precise confirmation has been made [33] by looking at the spectral lines of hydrogen from Sirius B and 40 Eridani B, a couple of bright white dwarf stars. Both of these stars are in binary star systems which makes a determination of the mass possible from the orbital parameters. The radius may be inferred from stellar models and the observed spectra. White dwarf stars have a mass similar to that of the Sun but are much more compact, having a radius about 100 times smaller. This increases the magnitude of the effect but unfortunately the spectra from these particular stars are rather diffuse and accurate measurements are not possible.

The most accurate data on the gravitational redshift comes from laboratory measurements utilizing the Mössbauer effect. In 1962, Pound and Rebka [34] placed a ^{57}Fe gamma ray source at ground level and detected the photons at the top of a 22.6 m tower at Harvard university. A more precise version by Pound and Snider [35], performed in 1964, agreed with the theoretical prediction to within 1 percent.

Exercise 15.5 *Compute the magnitude of the redshift in the Pound-Rebka experiment. Begin by showing that near the Earth's surface the magnitude of the redshift is*

$$z_g = \frac{GM}{Rc^2}\Delta h\,. \tag{15.21}$$

The gravitational redshift may be interpreted as due to the loss of energy light experiences as it climbs out of a gravitational field. (See problem???).

It is worth pointing out that the effect is often erroneously described in terms of a doppler shift. This is quite wrong as the source and detector have a fixed separation in the gravitational field. Any apparent motion noted by an observer in free-fall is due to non-local tidal effects.

15.2.2 Lengths in a gravitational field

It is tempting to suppose that lengths, as measured by a non-inertial observer in a gravitational field, will be contracted, just as times are dilated. However, length contraction stems from the relativity of simultaneity and the non-inertial observer simply doesn't have a set of synchronised clocks with which to measure the lengths of moving objects. In fact, we have just seen that standard clocks placed at different locations in the non-inertial frame will run at different rates. Therefore, it makes no sense to import a (freely falling) local inertial observer's point of view without considering how one should define length in a non-inertial frame. This we leave until later.

15.3 Curved Spacetime

Einstein's genius did not content itself with the Equivalence Principle. He proceeded to the conclusion that it meant spacetime in the vicinity of massive objects was curved. Gravitational effects were due to this curvature. Einstein called his theory the General Theory of Relativity. Most scientists nowadays prefer to call it Einstein's Theory of Gravitation, though Einstein was definitely motivated by a desire to escape from the strictures of inertial frames.

15.3.1 Non-Euclidean geometry

Euclidean geometry is the geometry of flat space. It was actually not until the 19th century that it was properly recognized that geometries could exist that did not satisfy all of Euclid's postulates. The discoveries of non-Euclidean geometries by Lobachevsky and Riemann (with important work also by Gauss and Bolyai) were the culmination of centuries of investigation of Euclid's fifth postulate concerning parallel lines.

In Euclidean geometry, parallel lines never intersect. This need not be the case though in non-Euclidean geometries and the behavior of parallel lines serves to distinguish those geometries. There are also other ways to characterize geometries. One is originally due to the Jesuit priest Saccheri in the 17th century. It is a well-known theorem of Euclid's that the sum of the interior angles of a triangle is $180°$. However, consider the surface of the Earth, which is essentially spherical. If we consider the triangle formed by the equator and two lines of longitude (meeting at a pole) then the two interior angles at the equator are each $90°$ and so the additional contribution of the interior angle at the pole clearly leads to a sum in excess of $180°$. Thus, one can classify the geometry according as to whether the sum of the interior angles of a triangle is less than, equal to, or greater than $180°$. Geometries in which the sum is less than $180°$ correspond to saddle-like surfaces.

Another way to classify the geometry is by comparing the circumference of a circle with its diameter. In Euclidean geometry the ratio of circumference to diameter is $\pi \approx 3.141592654$. However, consider again the surface of the Earth. Take as a circle the equator. A meridian passing through a pole provides a diameter and clearly, for a sphere, the length of this diameter is just one half the distance around the equator. Thus the ratio of circumference to diameter for this geometry is 2, which is less than π. Similarly, if we considered a circle on a saddle, then the diameter is relatively short and the ratio of circumference to diameter is greater than π.

15.3.2 Gravity as a curvature of spacetime

How then does one conclude that gravity can be interpreted as a curvature of spacetime? A simple example can help.

Suppose that one has a spinning disk. Then an observer stationed on that disk is non-inertial. That observer will experience imagined centrifugal forces trying to throw him away from the rotation axis. In order to calibrate his measuring rods he cannot simply follow the procedures we used earlier, involving the timing of light pulses because there is no guarantee that these procedures are valid in a non-inertial frame. In fact this observer cannot expect any of the physics developed for inertial observers to be valid in his frame. He can though compare with an inertial observer that is not on the disk and can import length and time standards from him. For example, if the observer is stationed on the edge of the disk, then, as seen by an inertial observer at rest with respect to the disk center, he has an instantaneous velocity given by Eq. 13.12. If the inertial observer arranges for a long rod to travel along a tangent to the disk with this same speed then he can mark off on this moving rod, the ends of the non-inertial observer's standard as it swings by (it being momentarily at rest with respect to the uniformly moving rod). But a rod moving uniformly in an inertial frame exhibits length contraction and so the inertial observer will determine that the non-inertial observer's standard is also short and will communicate this to him. The next step is tricky. One might think that the inertial observer would deduce that the non-inertial observer will measure a total circumference that is also shorter than for a non-spinning disk. However, one must remember that length contraction is due to the relativity of simultaneity. In measuring the circumference of a moving circle, one cannot simply mark off both ends simultaneously because the circumference is a length bent around on itself. In order to measure the circumference of a rotating circle one must mark off a point on the circumference and wait for it to come back to the same position. Since the inertial observer can measure the tangential speed, v, of the rim he would measure the circumference (in his frame) as

$$C = vT \tag{15.22}$$

where T is the measured period of rotation, in the inertial frame. This will coincide with the length he would obtain by laying out a stationary tape measure around the disk and is consistent with $C = \pi D$ since $v = r\omega$ where $\omega = 2\pi/T$ by definition. On the other hand, since a clock in the non-inertial frame is moving past at the speed v, the period indicated by a clock on the rim of the rotating disk will be dilated:

$$T' = \gamma T \tag{15.23}$$

and an inertial observer moving at this same speed will deduce that the circumference of the rotating disk is

$$C' = vT'. \tag{15.24}$$

This corresponds to having a tape measure sitting on the rim of the disk and being drawn away, without slipping, by the inertial observer (who is moving with respect to the disk center), as the disk rolls past in his frame. The tape measure is stationary in this observer's frame and the disk rolls by with speed v, completing one revolution in time T' according to synchronized clocks. But by comparison with the clocks in the frame of the disk center, T' is dilated, as above. Thus this observer will conclude that the rolling disk has a larger circumference than one at rest in his frame, i.e. one for which $C = \pi D$. The non-inertial observer on the disk has no choice but to accept the same conclusion. (Since the tape measure is stationary in this inertial observer's frame and the disk rolls by without slipping, one does not have to mark the points corresponding to one revolution simultaneously to obtain a valid length. Contrast this with an inertial observer in the frame of the disk center observing the same tape measure being drawn away. Not only can he not mark the points corresponding to one revolution simultaneously but the tape measure is moving. This observer cannot interpret any length of tape he might measure as the circumference of the disk.)

One now notes that the direction of a diameter is perpendicular to the instantaneous velocity of the rim and so undergoes no length contraction. Thus the ratio of circumference to diameter of the rotating disk is *more* than π and the non-inertial observer must conclude that he is living in a saddle-like geometry.

Similarly it is with the non-inertial observer on the surface of the Earth (or any other massive body). The time he measures is dilated in comparison to an inertial observer. Consequently, while the infinitesimal lengths he measures are contracted the larger lengths can only be defined via his time measurements and are longer. Hence he is forced to conclude that the geometry of spacetime, in his reference frame, is not the flat space of Minkowski spacetime. Qualitatively speaking, an observer on the surface of the Earth, deep in its gravitational field, is analogous to an observer on the rim of a spinning disk.

15.3.3 Retardation of light near massive bodies

Light in an inertial frame always travels with speed c. If the progress of light past a massive body is measured by a series of local inertial observers, then each will agree that light travels with speed c. However, in a gravitational field, each of the local inertial observers is accelerating with respect to the other — this being why they are only local. If we were to measure the effective speed of light, as inferred by an inertial observer far removed from the mass then we would find that it has been slowed down in passing the massive body.

15.3.4 Deflection of light near massive bodies

Because light is effectively slowed down in passing close to a massive object, it gets bent, just as light is refracted on passing into a medium where the refractive index is different from the one from which it exited.

Einstein predicted that starlight passing close to the Sun's rim would be deflected by 1.75" of arc. This is a factor of two greater than naive arguments using just the equivalence principle and this factor arises from the curvature of spacetime.

In order to see the deflection of starlight, one must wait for an eclipse. One can then see the stars near to the Sun and photograph them. The deflection is then determined by measuring the angular separation between stars close to the Sun and stars far from the Sun and comparing this with the same angular separations of these stars when the Sun is below the horizon (so that the stars are both visible and in a direction far from the Sun). This is a very difficult experiment to perform because it requires accurate measurements some months apart and the eclipse measurement must generally be done in some remote location where the eclipse just happens to occur.

An expedition led by Eddington, just after World War I, found the predicted deflection to within experimental errors and made Einstein an instant celebrity. The occasion was touted as the overthrow of Newton's gravitation law and proof that spacetime was curved. Actually though, the experimental errors were sufficiently large to render the experiment as not entirely conclusive. Further experiments of the same type have not led to significant improvement in accuracy.

Fortunately there is an alternative approach. One can use radio sources rather than light, which means that one doesn't have to wait for a solar eclipse but simply requires a radio source to pass behind the Sun. The quasar 3C279 in fact does this on October 8 each year. The method is made even more viable by the advent of long baseline radio interferometry which permits the angular separation of different sources to be measured to within 3×10^{-4} seconds of arc. Using this approach, Einstein's theoretical prediction has been confirmed to 1 part in 10,000.

Gravitational lensing

An extraordinary example of the deflection of starlight is afforded by very distant (and bright) objects called quasars, whose light just happens to pass close by a very massive nearer object such as a galaxy in the direct line of sight from Earth. Then, light may be bent sufficiently that it can reach Earth by passing on either side. Consequently, a pair of images of the distant quasar can be seen from Earth. Sometimes, the geometry permits more than two images and even arcs or rings. This effect is called gravitational lensing and several examples are known.

15.3.5 Precession of the perihelion

An early success of Einstein's theory was an explanation for the $43''$ per century precession of the perihelion of Mercury. In a curved spacetime a planet does not orbit the Sun in a static elliptical orbit, as in Newton's theory. Rather, the orbit is obliged to precess because of the curvature of spacetime. When Einstein calculated the magnitude of this effect for Mercury he got precisely the previously unexplained $43''$. He correctly took the view that this was an important confirmation of his theory.

Einstein's theory also correctly accounts for a smaller discrepancy of $8.6''$ per century in the precession of the perihelion of Venus. The value is smaller than that of Mercury because Venus is further from the Sun and the curvature of spacetime is less.

The binary pulsar PSR 1913+16

A pulsar is an extremely dense star, so dense that all of the empty space in an atom has been squezzed out and one is left with the individual atomic nuclei touching each other. Under these circumstances, the star is of the order of 10^{15} times denser than ordinary matter. In this state most of the protons in the nuclei have combined with atomic electrons to form neutrons and the star is therefore also known as a neutron star.

The binary pulsar PSR 1913+16 was discovered by Hulse and Taylor in 1974. This is a pair of neutron stars orbiting each other, with a semi-major axis of 2.3 light-secs. By measuring the orbital motion, one can deduce much information about the pulsars themselves.

An extraordinary feature of this binary system is that the extreme density and quite small orbital radius means that the orbital precession is an astounding $4.2°$ per year! The measured value agrees very well with Einstein's theory.

Indeed, this binary system has proven to be a rich testing ground for gravitational theories. One interesting aspect is that the period of the orbit is gradually speeding up as the pulsars spiral in towards each other. This implies a loss of energy — compare it to an artificial satellite gradually spiralling in to the Earth due to atmospheric drag — and the rate at which this is happening is consistent with the radiation of energy via gravitational waves — a vibration of the curvature of spacetime — in complete accord with the predictions of Einstein's theory.

15.3.6 Black holes

Black holes are a further prediction of Einstein's theory for which there seems to be some known candidates. A black hole is basically an object so dense, even denser than a neutron star, that spacetime around it is so curved that light can not escape from it.

Review

What is the equivalence priciple?

Does a clock run faster at high altitude or low altitude?

What is the gravitational redshift?

What is meant when one says that Newton's theory failed to explain the observed precession of the perihelion of Mercury? How big was the discrepancy and how is it explained in Einstein's theory?

Questions

1. Many elementary particles, such as the proton and the positron (i.e. the antiparticle of the electron), have the same electric charge. Why does this not mean that electric forces obey an equivalence principle?

Problems

1.

Bibliography

[1] Aristotle, *Physics*, Book IV. [reprinted in English in *Great Books of the Western World*.]

[2] Lucretius, *On the Nature of Things*.

[3] S. Clarke, *Recollections of papers which passed between Mr Leibnitz and Dr Clarke* (1717), p. 57

[4] Galileo Galilei, *Dialogue Concerning the Two Chief World Systems — Ptolemaic and Copernican*, 1632.

[5] Staff of the Space Dept., Johns Hopkins University Applied Physics Lab. and Staff of the Guidance and Control Lab., Stanford University, "A Satellite Freed of all but Gravitational Forces: 'TRIAD I' " J. Spacecraft, **11** 637-644 (1974);

[6] P. Melchior, *The Tides of the Planet Earth* (Pergamon, Oxford, 1978).

[7] John K. Fotheringham, *Historical Eclipses* (1921). D.R. Curott, "Earth deceleration from Ancient solar eclipses" Astron. J. **71**, 264-269 (1966).

[8] Euclid, *The thirteen books of Euclid's Elements* (circa 330-320 BC). [English translation from the text of Heiburg, with introduction and commentary by Sir Thomas L. Heath (Cambridge University Press, Cambridge, 1908); 2nd Ed. reprinted (Dover, New York, 1956)].

[9] Plato, *Timaeus*, in Great Books of the Western World (Encylcopedia Britannica).

[10] J.C. Hafele and R.E. Keating, “Around-the-World Atomic Clocks: Predicted Relativistic Time Gains” Science, **177**, 166-168 (1972); “Around-the-World Atomic Clocks: Observed Relativistic Time Gains” Science, **177**, 168-170 (1972).

[11] Sir Isaac Newton, “Mathematical Principles of Natural Philosophy”, 1686 [revised translation into English by Florian Cajori (Univ. California Press, Berkeley, 1934)].

[12] A.A. Michelson and E.W. Morley, “On the Relative Motion of the Earth and the Luminiferous Ether” Am. J. Sci. (3rd series) **34** 333-345 (1887).

[13] G. Joos, “Die Jenaer Wiederholung des Michelsonversuchs” Annalen der Physik **7** 385-407 (1930);

[14] R.S. Shankland, S.W. McCuskey, F.C. Leone and G. Kuerti, “New Analysis of the Interferometric Observations of Dayton C. Miller” Rev. Mod. Phys. **27** 167-178 (1955).

[15] R.J. Kennedy and E.M. Thorndike, “Experimental Establishment of the Relativity of Time” Phys. Rev. **42** 400-418 (1932);

[16] T.S. Jaseja, A. Javan, J. Murray and C.H. Townes, “Test of Special Relativity or of the Isotropy of Space by Use of Infrared Masers” Phys. Rev. **133A** 1221-1225 (1964);

[17] A. Brillet and J.L. Hall, “Improved Laser Test of the Isotropy of Space” Phys. Rev. Lett. **42** 549-552 (1979);

[18] Particle Data Group booklet, 1994.

[19] T. Alväger, F.J.M. Farley, J. Kjellman and I Wallin, “Test of the Second Postulate of Relativity in the GeV Region” Phys. Lett. **12** 260-262 (1964);

[20] K. Brecher, “Is the Speed of Light Independent of the Velocity of the Source?” Phys. Rev. Lett. **39** 1051-1054, 1236(E) (1977).

[21] H.E. Ives and G.R. Stilwell, “An Experimental Study of the Rate of a Moving Atomic Clock” J. Opt. Soc. Am. **28** 215-226 (1938); “An Experimental Study of the Rate of a Moving Atomic Clock. II” J. Opt. Soc. Am. **31** 369-374 (1941).

[22] M. Kaivola, O. Poulsen, E. Riis and S.A. Lee, "Measurement of the Relativistic Doppler Shift in Neon" Phys. Rev. Lett. **54** 255-258 (1985).

[23] D.H. Frisch and J.B. Smith, "Measurement of the Relativistic Time Dilation Using μ-Mesons" Am. J. Phys. **31** 342-355 (1963).

[24] J. Bailey, K. Borer, F. Combley, H. Drumm, F. Krienen, F. Lange, E. Picasso, W. von Rüden, F.J.M. Farley, J.H. Field, W. Flegel and P.M. Hattersley, "Measurements of relativistic time dilation for positive and negative muons in a circular orbit" Nature **268** 301-305 (1977).

[25] A.S. Goldhaber and M.M. Nieto, "The Mass of the Photon" Scientific American, (May 1976) p. 86 .

[26] See any book on mathematical physics, covering the subject of vector analysis, e.g. Mary L. Boas, "Mathematical Methods in the Physical Sciences" 2nd Ed. (Wiley, New York, 1966, 1983).

[27] R.H. Dicke, "Experimental Relativity" in *Relativity, Groups, and Topology* ed. C. DeWitt (Gordon and Breach, New York, 1964); P.G. Roll, R. Krotkov and R.H. Dicke, Ann. Phys. (N.Y.) **26** 442 (1964).

[28] V.B. Braginsky and V.I. Panov, Zh. Eksp. and Teor. Fiz. **61**, 873 (1971) [transl. in Sov. Phys. JETP **34**, 463 (1972)].

[29] J. Müller, M. Schneider, M. Soffel and H. Ruder, Astrophys. J. Lett. **382**, L101 (1991).

[30] C.O. Alley, "Relativity and Clocks" in Proceedings of the 33rd Annual Symposium on Frequency Control (Electronic Industries Association, Washington, 1979).

[31] R.F.C. Vessot, M.W. Levine, E.M. Mattison, E.L. Blomberg, T.E. Hoffman, G.U. Nystrom, B.F. Farrel, R. Decher, P.B. Eby, C.R. Baugher, J.W. Watts, D.L. Teuber and F.D. Wills, "Test of Relativistic Gravitation with a Space-Borne Hydrogen Maser" Phys. Rev. Lett. **45** 2081-2084 (1980).

[32] J.W. Brault, Princeton University thesis, 1962 [see also R.H. Dicke, in C. deWitt and B. deWitt, eds., *Relativity, Groups and Topology* (Gordon and Breach, New York, 1964) p. 170; J.L. Snider, Phys. Rev. Lett. **28**, 853 (1972).

[33] J.L. Greenstein et al., Astrophys. J. **169**, 563 (1971).

[34] R.V. Pound and S.A. Rebka, “Apparent weight of photons” Phys. Rev. Lett. **4** 337-341 (1960).

[35] R.V. Pound and J.L. Snider, Phys. Rev. Lett. **13**, 539 (1964); Phys. Rev. B **140**, 788 (1965).

Answers to Exercises

2.1 11.8 in