Thomas Strohm

Special Relativity for the Enthusiast

Springer

Special Relativity for the Enthusiast

Thomas Strohm

Special Relativity for the Enthusiast

Springer

Thomas Strohm
Asperg, Germany

ISBN 978-3-031-21923-8 ISBN 978-3-031-21924-5 (eBook)
https://doi.org/10.1007/978-3-031-21924-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*This book is dedicated to my beloved family:
Yodaira, Lena and Salma Sofia.*

Preface

The **special theory of relativity** is a fascinating topic. It predicts effects that are blatantly counterintuitive and which have puzzled people since Einstein created the theory in 1905. Relativity contains a wealth of interesting physics for a cheap price: the mathematics needed to understand this physics is relatively easy. And relativity is a very well-tested pillar of modern physics. Without Einstein's theory, we would not be able to understand large parts of physics, not to mention important technologies like the Global Satellite Navigation System.

Our approach. Consequently, there are many books featuring an introducing to the special theory of relativity. Most of them put too much emphasis on mathematical formulation. To understand the physics behind the theory, however, is much more fruitful. Therefore, this book goes a long way towards making the reader understand the physics of this wonderful theory. The book also includes the historical perspective wherever it eases understanding and explains the situation before Einstein cut the Gordian knot in 1905.

There are several things that we do differently from most other authors. For instance, you will not find the concept of rest mass in this book. By the **mass of an object**, we always mean the mass m of the object at rest, and therefore, we can do without the attribute "rest". The "relativistic mass" $m_{\text{rel}} = \gamma m$ (or "mass of a moving body") will not be used. There are several reasons for this. First, because of $E = m_{\text{rel}}c^2 = \gamma mc^2$, the relativistic mass would be nothing but the energy of the object. Occam's razor finds it dispensable and tosses the relativistic mass on the garbage heap of relativistic physics. Thus, the relativistic mass would not transform as a tensor in Lorentz transformations and therefore has no place in a modern covariant formulation of the special theory of relativity. This, eventually, was also Einstein's point of view.[1]

[1] Einstein himself writes (the author's translation): "It is not good to speak of the mass $m_{\text{rel}} = \gamma m$ of a moving body, since no clear definition can be given for m_{rel}. It is better to restrict oneself to the rest mass m. In addition, one can use the expression of momentum and energy if one wants to indicate the inertial behaviour of rapidly moving bodies." [in a letter from Einstein to L. Barnett, from 19.06.1948, cited in L. B. Okun, Physics Today **43**, 31 (1989)].

Our access to **energy and momentum** in the special theory of relativity is via the conservation laws.[2] In this vein, in Sect. 13.1, we present Einstein's original derivation of the relativistic energy, and in Sect. 13.3, the derivation of the relativistic momentum given by Lewis and Tolman in 1909 [LewisTolman09]. These derivations are relatively easy, conceptually clear and carried out by discussing carefully crafted *Gedanken* experiments.

Eventually, we will require a word about the adjective "relativistic". A theory is relativistic if it obeys the principle of relativity. This is true for both, classical mechanics (via the Galilei transformation) and Einstein's special theory of relativity (via the Lorentz transformation). Only Einstein's theory, however, is also correct for large velocities, as well as electromagnetic phenomena. Therefore, if we occasionally write "relativistic mechanics", we indeed refer to Einstein's theory.

Structure and target group. This book is not intended for a linear read. It can be seen as consisting of two large parts. The first part, up to and including Chap. 11, is comparatively easy and does not require a lot of previous knowledge in physics and mathematics.

The second part, starting around Chap. 12 *en large*, is more challenging and requires solid college physics and mathematics.

In addition to that, the whole part is interwoven with sections marked as **digressions**. In these sections, we dig deeper, addressing material that is more sophisticated and sometimes using concepts that are only introduced in the basic text later. For beginners, it makes sense to first read the book from beginning to end without the *digression*s and then go back and read the latter.

The book covers all traditional material on special relativity, which is mostly about a century old, but there are also **several modern sections**, in particular, on the measurement of the speed of light (Sect. 6.3.5), on atomic clocks (Sect. 9.4) and on satellite navigation (Sect. 9.11.3).

An important limitation of special relativity is that it does not cover gravitation. And there is no easy way to extend it. Einstein had to develop a considerably more comprehensive theory, the **general theory of relativity**, to include gravitation. For the case of small gravitational fields, general relativity becomes special relativity, so learning special relativity is not a waste of time. On the contrary, learning general relativity requires a mastery of special relativity. In Chap. 15, the last chapter of this book, we sketch the first steps towards general relativity.

Recommended literature. There are many expositions of the special theory of relativity. Most textbook series on theoretical physics include a chapter on relativity. Usually, the presentations are very theoretical and concentrated and therefore not suitable for obtaining a deeper understanding of relativity. Examples are as follows:

- R. P. Feynman, The Feynman Lectures, Vol. 1, Chap. 15 (Addison-Wesley Longman, 1970).

[2] Alternatives would be the ugly Ansatz $p_{\mathrm{rel}} = m_{\mathrm{rel}} v$ for the relativistic momentum p_{r} and its manifest generalization $p_{\mathrm{rel}} = mu$, where u is the relativistic velocity, which is not very intuitive. Also, the relativistic generalization of the force is surprisingly subtle (see, e.g. Thorsten Fließbach, "Die relativistische Masse" (Springer 2018), which unfortunately is available only in German).

- L. D. Landau and E. M. Lifshitz, “The Classical Theory of Fields”, Chaps. 1–3 (Pergamon Press, 1971).
- W. Nolting, “Theoretical Physics 4: Special Theory of Relativity” (Springer, 2017).
- and many more.

Special expositions of special relativity that are particularly recommendable are as follows:

- N. D. Mermin, “It’s about time” (Princeton University Press, 2009). A very nice book that uses a geometrical approach similar to that used in our book.
- W. Rindler, “Relativity” (Oxford University Press, 2006). Very concise, and well explained.
- B. Russell, “ABC of relativity” (Allen & Unwin, 1925). This is very fundamental, with a penchant for the philosophical.
- W. Pauli, “Theory of relativity” (Dover, 1958). A reprint of the famous Mathematical Encyclopedia article.
- E. F. Taylor and J. A. Wheeler, “Spacetime physics” (Freeman and Company, 1992).
- A. P. French, “Special relativity” (W W Norton & Company, 1968). A very well-written book with a good selection of material and profound explanations. In concept, this book is similar to ours.
- G. Barton, “Introduction to the Relativity Principle” (Wiley, 1999). A very complete book with a considerable amount of detail.

Concepts, Terms, Notation. A few comments are regarding the **nomenclature** used. For events, you will find both notations, (x, t) and (t, x). In different books, you will find different notations, each with its own advantages and disadvantages. Instead of trying to unify them, it is better to get used to both.

By **classical mechanics**, we mean mechanics before the advent of special relativity, which uses the Galilei transformation when changing inertial frames. **Classical physics** denotes classical mechanics and electrodynamics, but with the luminiferous aether. Eventually, we will use the term **modern physics** for the physics that emerged from the modernization of classical physics by special relativity and, occasionally, quantum physics.

Explaining the effect of **aberration** in classical physics is a challenge. And that’s not really because we have to consider the medium (luminiferous aether). The fact that the phase velocity and the group velocity transfer in different ways when changing from one (inertial) reference frame to another is difficult to deal with. Furthermore, there is a dangerous trap: the similar-looking Doppler effect is an effect between the wave’s source and the observer, while aberration is an effect between two different observers.

With these words, we want to liberate the reader and hope that he/she enjoys this book. This will help a lot in understanding the wonderful physics of special relativity.

Asperg, Germany Thomas Strohm
March 2022

Contents

Chapter 1
Introduction

Measured by the standards of the classical physics of around 1900, the special theory of relativity is counterintuitive and, in this perspective, revolutionary.[1]

We will demonstrate that **classical mechanics** (or *Newtonian mechanics*), as you learned it, fails if very large velocities are involved. It fails because its predictions become false. When explaining the motion of particles in particle accelerators, classical mechanics is no longer applicable. We will see this in Chap. 3 in the discussion of a particular experiment. As a consequence, we have to replace classical mechanics with a new theory, **relativistic mechanics**. This relativistic mechanics was worked out by Albert Einstein in the context of his **special theory of relativity (SR)** (or **special relativity**). His achievement was made "on the shoulders" of colleagues like Hendrik Antoon Lorentz, Henri Poincaré, and others who laid substantial groundwork. But only Einstein was able to cut a Gordian knot and reshape a very complicated and intricate theory so that it would take the crystal clear form of the special theory of relativity. The book at hand is about this **relativistic physics**.[2]

Starting with Chap. 7, we will develop Einstein's special theory of relativity, which was published in a seminal paper in the Annalen der Physik under the title "Zur Elektrodynamik bewegter Körper" [Einstein05a].[3]

The need for Einstein's theory can be seen rather easily. Imagine Alice and Bob (see Fig. 1.1). Alice has a lamp that emits a light beam (or a light pulse). Bob has a spaceship and flies by Alice and the lamp with a large velocity v. Then, both Alice and Bob independently measure the velocity of the light beam (using the same type

[1] This is also the same time when quantum theory was in the making, an even weirder idea.

[2] Relativity also plays an important role in classical mechanics. We will discuss Galilei's principle of relativity in Chap. 3. In special relativity, however, the principle of relativity has much more power. Therefore, the literature sometimes refers to classical mechanics with the term *non-relativistic mechanics*—although it is, in a sense, a relativistic theory.

[3] This paper is freely available on the Internet. Have a look at it!.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_1

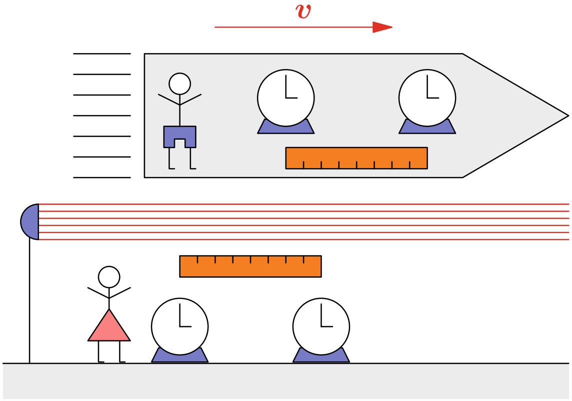


Fig. 1.1 Alice measures the speed of a light beam emitted by a lamp next to her. Bob is in his spaceship, which rapidly flies by Alice and the lamp. He also measures the speed of light of the same light beam emitted by the lamp

of measurement device). Believe it or not, both get the same velocity, the **speed**[4] **of light** $c = 299{,}792{,}458\,\mathrm{m/s}$ (sometimes referred to as *light speed*). How is this possible? If you aren't puzzled by this result, think again as to what Alice and Bob should get as measurement results.

By the way: the fact that the speed of light is "absolute" (independent of the observer) is built into the Global Positioning System (GPS). If the speed of light depended on the observer, that is, if Alice and Bob measured different velocities, navigation with GPS would not be possible (see Sect. 9.11.3). Note that the speed of light is huge. To travel the distance from the Earth to the Moon, a light pulse needs only a little bit more than a second!

Our argument above involving Alice and Bob is what is know as a *Gedanken* experiment. In Sect. 2.1, we discuss an experiment that actually has been conducted and that also shows that classical mechanics is no longer applicable for very large velocities.

[4] In English, the term *velocity* means a direction and a magnitude, e. g., "5 km/h to the west". The term *speed*, however, refers only to the magnitude of a velocity. We will use the term *speed* only in the established term "speed of light" (and the like) and will otherwise talk about the *magnitude of the velocity*.

Chapter 2
The Limits of Classical Mechanics

No doubt, you are used to the central elements of **classical mechanics** and know that it describes many aspects of Nature very well. Adjacent to classical mechanics is **electrodynamics**, which describes electromagnetic phenomena. **Light** is a phenomenon of **wave optics**, a special case of electrodynamics.

In this chapter, we will show that there are mechanical phenomena that classical mechanics cannot describe correctly. We will shortly present a very instructive experiment. It shows that **classical mechanics is no longer applicable when "large" velocities are involved**.

By **small velocities** v, in this book, we denote velocities that are *much smaller* than the speed of light c. For such velocities, $v \ll c$ or $v/c \ll 1$ holds.

We will find out later that matter cannot be moved faster than the speed of light.[1] And by **large velocities**, we mean velocities that are in the order of the speed of light. This, depending on the situation, can be 1% of the speed of light or more. Remember that the speed of light is huge! What we understand to be large velocities in our everyday world (for instance, the velocity of a rocket) would be, according to our definition, significantly "small" velocities indeed.

Now, we present the promised famous experiment that clearly shows that classical mechanics no longer applies when large velocities are involved. It yields incorrect results, results that fail to describe the outcome of experiments like that put forth in the next section.

[1] This formulation is a little bit sloppy. In Chaps. 7.8.1 and 13.1.3, we will look at these issues in detail.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_2

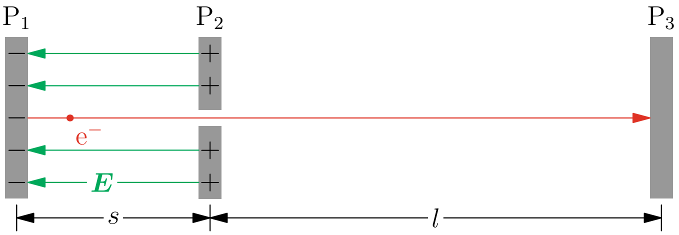


Fig. 2.1 Principle of the Bertozzi experiment

2.1 The Bertozzi Experiment

2.1.1 Introduction

Classical mechanics tells us that an object can be brought (accelerated) to arbitrary large velocities. One just needs a force that acts on the object for a sufficiently long amount of time. Take an object of mass m that is initially at rest and upon which a force F_0 acts for a period Δt of time. After this period, the object has the velocity $v = a\Delta t = (F_0/m)\Delta t$, where a is the acceleration. According to classical mechanics, one can achieve arbitrary velocities by making the period of acceleration Δt sufficiently large. The formula shows: the larger the force, the smaller the mass, and the longer the acceleration, the better for the final velocity.

Classical mechanics also tells us that the kinetic energy E_{kin} of the object is quadratic in the velocity v: $E_{\mathrm{kin}} = mv^2/2$. This relation between energy and velocity has been checked for large velocities in several experiments. The first experiments in this direction were carried out by Walter Kaufmann and others starting in 1901, several years before Einstein developed his special theory of relativity. A more modern and, for didactic purposes, more suitable experiment for checking the formula for the kinetic energy was performed by William Bertozzi in 1964 at MIT in Boston [Bertozzi64].[2] This is the experiment that we present now.

2.1.2 Principle

The principle of the Bertozzi experiment is pretty easy (see Fig. 2.1). Exactly as described above, an object is accelerated. In the Bertozzi experiment, the object is an electron and we denote its mass by m_{e}. This electron is located in a homogeneous electric field $|\mathbf{E}_0|$ of an accelerator between two charged plates P_1 and P_2, which are separated by the distance s. Between the plates, there is a voltage U, which

[2] At http://education.jlab.org/scienceseries/ultimate_speed.html, there is a small film that you should have a look at.

implies an electric field of magnitude $|\mathbf{E}_0| = U/s$ in the space between the plates. Therefore, a force $F_0 = q_e|\mathbf{E}_0|$ acts on the electron, where $q_e = -e_0$ is the charge of the electron and $e_0 > 0$ the elementary charge. The electron is created at the left plate P_1 (extracted from the plate) and accelerated approximately from rest until it arrives at plate P_2. In the center of plate P_2, there is a small hole. The electron flies through this hole[3] and, when it passes, the time of flight t_0 is measured. Then, the electron continues its motion, now without any acting force, and after a distance l at time t_1, it hits another plate P_3. Between the plates P_2 and P_3, the velocity of the electron is constant and given by $v = l/(t_1 - t_0)$.

The acceleration voltages used by Bertozzi were on the order of 1 MV, causing a very large force to be applied to the electron. Together with the small mass of the electron, this implied huge accelerations on the order of 10^{16} times the gravitational acceleration at the surface of the Earth!

As mentioned, we want to check the formula for the kinetic energy. The mass of the electron and its velocity (after the acceleration phase) are known to us. What about the kinetic energy?

Bertozzi has determined the electron's kinetic energy with two different methods. The first method is based on the known acceleration voltage U. With $W = q_e U$, we can calculate the work performed by the electric field on the electron. Because the electron was at rest at the beginning, this work is equal to the kinetic energy of the electron at the end of the acceleration phase.

The second method makes use of the fact that the plate P_3 heats up when the electron hits it. Then, the electron transfers its kinetic energy in the form of heat to the plate. This causes the plate to experience a temperature increase, which can be measured.[4] Using the heat capacity of the plate, the transferred energy can be calculated. The comparison of both methods gave the same results, up to a small measurement error. In this way, the kinetic energy of the electron is determined and the formula for the kinetic energy can be checked.

2.1.3 More Details

The devil is in the details, and the experiment is a little bit more difficult as described above (see Fig. 2.2). First, it is not carried out with a single electron, but with a small cloud of electrons. This cloud is produced and accelerated in a van de Graff generator. The production happens with an electron gun (in the figure on the left) and the acceleration with a series of charged plates with a total voltage of 0.5 MV, 1.0 MV or 1.5 MV. When the electrons leave the van de Graff generator, they enter a linear accelerator ("linac"), which is an 8.4 m long evacuated tube. With the linear accelerator, the electrons can be further accelerated. At the entrance of the linear accelerator, the electrons fly through a small metallic tube (corresponding to the

[3] We won't consider electrons that hit plate P_2.

[4] This is not possible with one electron; there need to be a lot of electrons hitting the plate.

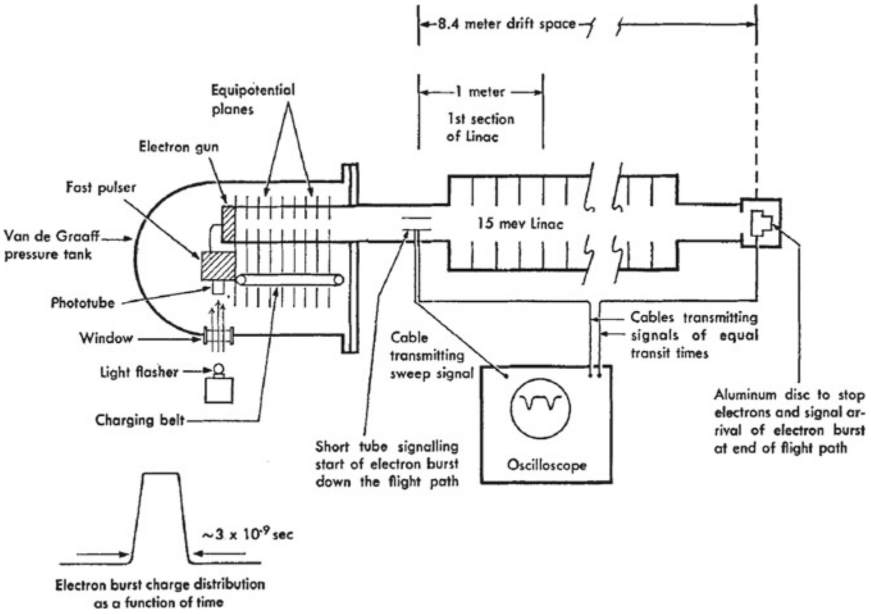

8.4 meter drift space
1 meter
1st section of Linac
Equipotential planes
Electron gun
Fast pulser
Van de Graaff pressure tank
Phototube
Window
Light flasher
Charging belt
15 mev Linac
Cable transmitting sweep signal
Cables transmitting signals of equal transit times
Aluminum disc to stop electrons and signal arrival of electron burst at end of flight path
Oscilloscope
Short tube signalling start of electron burst down the flight path
~3 x 10^{-9} sec
Electron burst charge distribution as a function of time

Fig. 2.2 Composition of the Bertozzi experiment (from Bertozzi's original publication)

plate P_2). Some electrons stick there and produce a small voltage signal. This signal is used to determine the time t_0 when the cloud enters the linear accelerator. At the end of the linear accelerator, 8.4 m after the small tube, the electrons hit a small aluminum disk, where they are stopped. Here, the time t_1 is determined, again, by a small voltage signal caused by the charging of the aluminum disk. The time difference, together with the distance traveled yields the final velocity of the electrons.

If the electrons experience a further acceleration in the linac, this has to be taken into account when the final velocity is calculated (see Exercise 1). The energy of the electrons as described above was determined with two different methods. In the second method, the temperature increase of the aluminum disk when the electrons hit it is measured. To be able to calculate the kinetic energy of one electron, the number of electrons in the cloud has to be determined. This can be done by measuring the charge deposited on the disk by the electrons.

2.1.4 Result

The result of the experiment is shown in Fig. 2.3. Instead of displaying the measured kinetic energy E_{kin} as a function of the measured velocity v, Bertozzi has drawn the quantity v^2/c^2 as a function of $E_{\text{kin}}/m_{\text{e}}c^2$, i.e., the square of the measured velocity v in units of the speed of light versus the energy transferred to the electron in the acceleration phase in units of the constant $m_{\text{e}}c^2$. This looks complicated, but actually is very useful. Both axes then are without units (comparing this with

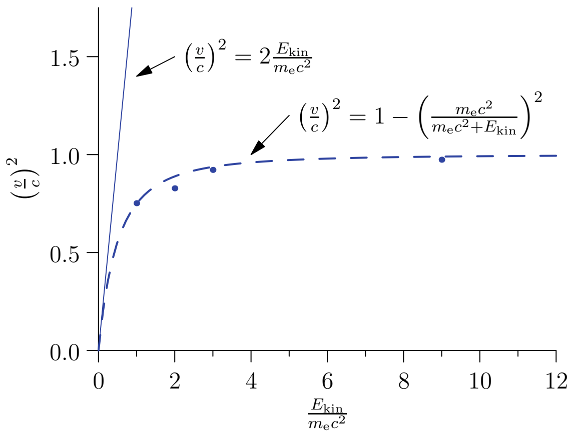


Fig. 2.3 Result of the Bertozzi experiment. E_{kin} is the kinetic energy. The straight blue line is the prediction of classical physics and the blue dashed line is what is expected according to special relativity. The blue dots are Bertozzi's measurement results

the formula for the kinetic energy shows immediately that $m_e c^2$ has the unit of an energy). According to classical mechanics, one expects $E_{\text{kin}} = m_e v^2/2$, which is equal to $(v/c)^2 = 2 \cdot E_{\text{kin}}/m_e c^2$. This is a line through the origin with slope 2.

The points in the diagram represent the measurement values (the fifth measurement value with $E_{\text{kin}}/m_e c^2 = 30$ is far outside of the diagram). Obviously, for large velocities, the points lie far from the expectations given by the steep line. Classical mechanics fails. According to Einstein's theory, for the relation between the kinetic energy and the velocity (which we will derive in Sect. 13.1), one expects the dashed line, which fits the measurement result very well.

A look at the diagram shows that the electrons do not become faster than light. Bertozzi's experiment cannot prove this, but does indicate it. In Sect. 7.8, we will see that the special theory of relativity requires this: electrons can never move faster than light (this also holds for other objects).

2.1.5 Discussion

Though the electrons in Bertozzi's experiment, according to classical mechanics, should move much faster than light, this is never observed. For large velocities, classical mechanics makes completely incorrect predictions: neither do the electrons become arbitrarily fast nor does the relation $E_{\text{kin}} = mv^2/2$ hold for large velocities anymore. Classical mechanics cannot be applied if large velocities are involved.

Consequence of Bertozzi's experiment: Classical mechanics is no longer valid if large velocities are involved.

Note that, apart from Bertozzi's experiment, there are many other experiments that also demonstrate that classical mechanics fails for large velocities. The excuse that classical mechanics is fine but there was something wrong in Bertozzi's experiment is definitely not tenable.

Now, one could still argue that only the expression $E_{\text{kin}} = (m/2)v^2$ of classical mechanics is wrong, but not classical mechanics as a whole. This position is also not sustainable, because it is possible to discuss the experiment without using the concept of energy, but only the force.

To rebut a further objection, we remark that (non-relativistic) quantum theory does not come to a rescue either. The fact that an electron cannot move faster than light doesn't have anything to do with it being a quantum object and not a particle.

For more than two centuries, classical mechanics was the measure of all things in physics. But at the beginning of the 20th century, the limits of classical mechanics were demonstrated twice. The first time was by quantum theory, which shows that classical mechanics fails in the world of very small dimensions. Around the same time, Einstein's relativity showed that classical mechanics is not valid if large velocities play a role (or large gravitational fields). But this does not derogate the usefulness of classical mechanics. This theory (as do all others as well) has a particular limited scope of application and, under this restriction, describes the physics perfectly. A physical theory is not right or wrong; it is useful or not. And it is not useful if it gives incorrect predictions. Moreover, a physical theory is always only an approximation of Nature (although it may be a very precise approximation).

Thus, classical mechanics is not usable for large velocities. In the next chapters, we will investigate in detail how this failure comes about and see that it is necessary to replace classical mechanics with a new *relativistic mechanics*. This, as already mentioned, was achieved by Einstein in 1905 with his special theory of relativity. We will follow his footsteps.

Exercise 1: An electron is accelerated from rest with a voltage of 1 MV. Calculate the velocity of the electron according to classical mechanics and compare it to the speed of light.

Exercise 2: In Fig. 2.3, the relativistically correct relation between the kinetic energy and the velocity of an object with mass m is shown:

$$\left(\frac{v}{c}\right)^2 = 1 - \left(\frac{mc^2}{mc^2 + E_{\text{kin}}}\right)^2 .$$

Explain why, according to this formula, the velocity of an object cannot become larger than the speed of light c.

Exercise 3: In Bertozzi's experiment, electrons are produced with an electron gun. Find out how such a device works.

Exercise 4: Why is it possible, just on the basis of the results of Bertozzi's experiment, to claim that classical mechanics no longer works for large velocities? Why isn't it possible to deduce that objects cannot be accelerated to velocities larger than the speed of light?

Chapter 3
The Relativity Principle of Classical Mechanics

In the last chapter, we have seen that classical mechanics no longer works if large velocities are involved. It must be replaced with a theory that is also able to explain experimental results if objects move very quickly. Beginning with Chap. 6, we will show that Einstein's special relativity fulfills this need. But before working out this theory bit by bit, we must lay the ground and understand some indispensable foundations of classical mechanics.

3.1 Reference Frames

If we measure positions, velocities or the like, we always have to state what we refer to. The speedometer in a car displays a velocity **relative** to (or *with respect to*) the street, this is obvious. In the case of an airplane, it is already a bit more complicated. Is the airplane's velocity meant to be relative to the Earth's surface or relative to the surrounding air?

The system, that we refer to, is the **reference frame**. In particular, the velocity that one measures always depends on the reference frame.

If a ship has a maximum velocity of 20 knots, this is meant relative to the reference frame in which the water is at rest. If we stand on the bank of a fast flowing river and measure the velocity of a ship that drives down the river *relative to the river bank*, we obviously can get a velocity which is larger than the maximum velocity of the ship because latter is meant to be understood relative to the resting water.

Exercise 5: How could an airplane measure its velocity relative to the Earth's surface or relative to the air that surrounds it? Discuss.

We imagine a reference frame that is always equipped with a **coordinate system**, and, in particular with a **cartesian** one: three coordinates x, y, and z whose axes are mutually *perpendicular* have the same *scale* and meet at the *origin* where all coordinates are zero. Suppose we have a rod 1 m in length and make it coincide with the x-axis. Then, the difference of the x-coordinates of both ends of the rod is 1 m.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_3

And if we make it coincide with another of the axes, the same holds. Alternatively, we can also place an arbitrary large sphere with its center at the origin of the coordinate system. Then, the sphere must intersect all the axes at the same coordinate.

The position of the origin of the coordinate system, as well as its orientation, can be freely chosen in the reference frame, a beneficial situation because it can often considerably ease the calculations. However, it is important that the coordinate system is at rest relative to the reference frame.

Many physical quantities have to be understood **relative to a particular reference frame**: velocity, kinetic energy, etc.[1]

One special reference frame is the **rest frame** of an observer or an object in which said observer or object is at rest. Another one is the **center-of-mass frame**, in which the center of mass of a system is at rest. The most important class of reference frames is that of **inertial frames**, which we will discuss in a minute.

Exercise 6: How do the Earth and the Moon move in their common center-of-mass frame? Where is the center of mass located?

3.2 Newton's Laws

Newton's laws, developed by Isaac Newton and published in 1687 in the *Philosophiae Naturalis Principia Mathematica*, form the core of classical mechanics. These laws deal with the motion of objects and of forces that act upon these objects. An important idealization is the **mass point**, which is an object without any extension whose only property is its *mass*. Working with such mass points, one does not have to concern oneself with rotations, internal energy, etc., and its position can be given by a single point in space, i. e., three numbers.

Newton's *first law* is the **law of inertia**. This law states that an object upon which no force acts, moves uniformly. **Uniform motion** means that the object moves along a (straight) line[2] with the magnitude of its velocity being constant (or it is at rest, which is the special case of moving with zero velocity). The law is valid only within the already mentioned inertial frame. We will come back to this in much more detail later.

Newton's *second law* is the **force law**, which relates changes of the motion state of objects to forces that act upon these objects. A force $\boldsymbol{F}$ acting upon an object causes an acceleration $\boldsymbol{a} = \boldsymbol{F}/m$ in it.

[1] Counterexamples are the mass and the charge.

[2] Note that a *line*, by definition, is always straight. Otherwise, it is a *curve*. Nevertheless, we will usually use the pleonasm "(straight) line", just to avoid any possible confusion.

Newton's *third law*, the **action-reaction law**, states that if two objects A and B are interacting and A exerts a force (action) on B, then B exerts a force (reaction) of the same magnitude, but in the opposite direction, to A. This law is nothing but the statement that momentum is conserved.

Newton's first and third laws are valid only in an inertial frame. The second law, however, can be modified by the addition of **fictitious forces** so as to also make it valid in some non-inertial frames.

3.3 Inertial Frames

3.3.1 Fictitious Forces

Some reference frames are better suited than others for the description of the motion of objects. Let us go to the fair and consider a chairoplane (see Fig. 3.1). Seats are attached to a rotating rim with chains hanging down. People sit in these seats. If the rim is at rest, the chains hang vertically. This is expected because of gravity. But if the rim rotates, the seats on their chains no longer hang vertically, but are slanted outwards. How do we explain this?

The **observer next to the chairoplane** (which rests relative to the Earth's surface) explains what she sees as follows (see Fig. 3.1, left side): the persons in the chairoplane move along a circle with a velocity of constant magnitude. For this to be possible, it is necessary to have a force that points to the center of the circle.

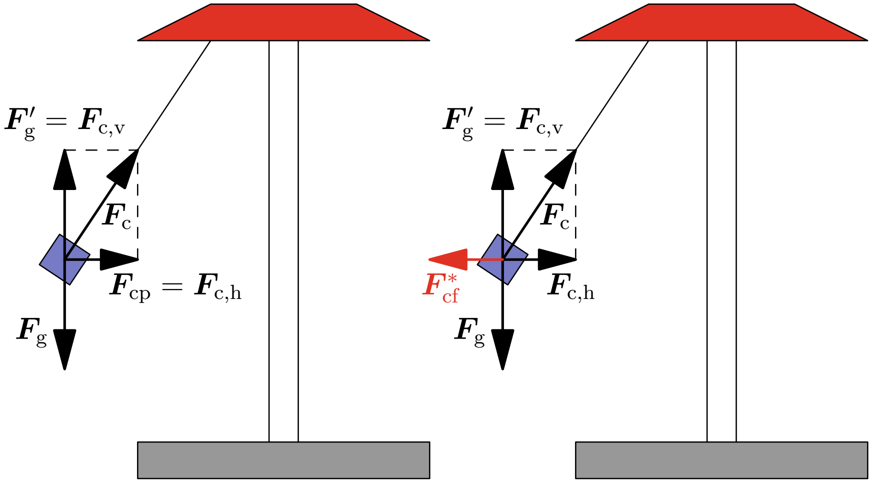

$\boldsymbol{F}'_{\mathrm{g}} = \boldsymbol{F}_{\mathrm{c,v}}$
$\boldsymbol{F}_{\mathrm{c}}$
$\boldsymbol{F}_{\mathrm{cp}} = \boldsymbol{F}_{\mathrm{c,h}}$
$\boldsymbol{F}_{\mathrm{g}}$
$\boldsymbol{F}'_{\mathrm{g}} = \boldsymbol{F}_{\mathrm{c,v}}$
$\boldsymbol{F}_{\mathrm{c}}$
$\boldsymbol{F}^*_{\mathrm{cf}}$
$\boldsymbol{F}_{\mathrm{c,h}}$
$\boldsymbol{F}_{\mathrm{g}}$

Fig. 3.1 Forces in a chairoplane. Left: For the observer at rest relative to the Earth's surface. Right: For the observer rotating with the chairoplane

This force is the **centripetal force**[3] $\boldsymbol{F}_{\mathrm{cp}}$. The centripetal force is an **abstract force** named after its function. It needs a real force that plays its role, and this force is given by the horizontal component $\boldsymbol{F}_{\mathrm{c,h}}$ of the chain force $\boldsymbol{F}_{\mathrm{c}}$, the force with which the chain pulls at the seat. The vertical component $\boldsymbol{F}_{\mathrm{c,v}}$ is the **counterforce** $\boldsymbol{F}'_{\mathrm{g}}$ to the gravitational force $\boldsymbol{F}_{\mathrm{g}}$ (without this counterforce, the seat would fall down to Earth). The angle α under which the chain deviates from the vertical adjusts such that the horizontal component of the chain force plays the role of the centripetal force. The larger this angle, the larger the horizontal component $\boldsymbol{F}_{\mathrm{c,h}}$ of the chain force. This agrees with our observations, because the faster the rotation, the larger the centripetal force must be. With the horizontal component of the chain force, there is a **real force** that causes the trajectory of the people in the chairoplane. We also feel this force when we hold an oscillating pendulum.

The **observer that rotates with the chairoplane** ("in the rotating reference frame") explains the effect in a different way (see Fig. 3.1, right side): there is a horizontal force that acts on the person in the seat, namely, the horizontal component $\boldsymbol{F}_{\mathrm{c,h}}$ of the chainforce $\boldsymbol{F}_{\mathrm{c}}$. The person should therefore move toward the rotation axis of the chairoplane, but this does not happen. For this reason, this force must be compensated by a counterforce, which is the **centrifugal force**[4] $\boldsymbol{F}^{*}_{\mathrm{cf}}$. The centrifugal force is a **fictitious force** that was invented simply and solely to explain the observations (this is indicated by the star in the notation of the force). A reason for this force cannot be given. Fictitious forces also violate Newton's third law. Fictitious forces do not have counterforces.[5]

Exercise 7: Consider an object on the rear window shelf of your car. If the car moves straight ahead, the object stays in its place. But if the car makes a curve, the object flies toward the exterior direction of the curve. How is this observation explained? Discuss this in the reference frame of the Earth (observer at the roadside) and in the reference frame of the car. In which reference frame are fictitious forces needed?

3.3.2 Inertial Frames

If no force acts on an object, it moves uniformly. This is Newton's law of inertia. In the cases above, we have seen that an object without an acting force has changed its state of motion. We remedied the problem by introducing a fictitious force. But the description in the reference frame without fictitious forces is always easier and clearer.

[3] From the Latin *centrum*, for "center", and *petere*, for "to strive for", "to seek". The centripetal force "strives for the center" of the particle's trajectory.

[4] From the Latin *fugare*, which means "to drive away" or "to chase away". The centrifugal force therefore is the force that chases an object away from the center.

[5] In the end, this is related to the fact that, in non-inertial frames, momentum conservation does not hold.

Reference frames in which the law of inertia holds, are called ***inertial frames*** (sometimes also *inertial reference frames* or *inertial systems*). In such reference frames, no fictitious forces are needed to describe the motion of objects. The observer that stands next to the chairoplane or the observer at the roadside are at rest in an inertial frame, while the observer rotating with the chairoplane or that sitting in the car are not.

Once we have found one inertial frame, we know all of them. Why? Take two inertial frames I and I'. An object upon which no force acts moves in both inertial frames uniformly (but with a different velocity). This is exactly the definition of an inertial frame. Thus, the inertial frames relative to each other also must move uniformly. The reverse conclusion also holds. Take an inertial frame I and a reference frame B that relative to I moves uniformly. An object upon which no force acts moves uniformly in I. But then, in B, the same holds, and therefore B is also an inertial frame. Eventually, we have the following conclusion: *Given an inertial frame I, all reference frames that move uniformly relative to I are also inertial frames, and this exhausts the set of inertial frames.*[6]

A car that moves accelerated on the Earth's surface is not an inertial frame. Is the Earth's surface an inertial frame? Almost, but not really. If it were an inertial frame, the oscillation plane of a Foucault pendulum would not rotate. To explain the Foucault pendulum and similar effects, we have to introduce fictitious forces: the centrifugal force and the Coriolis force. Due to the small rotational velocity of the Earth, these forces are also very small. For typical experiments in your physics course, the deviation of the Earth's surface from an inertial frame is insignificant: the Earth's surface in these cases is, to a very good approximation, an inertial frame.

Exercise 8: Inform yourself about the Coriolis force. How does it come about? Is it a "real" force or a fictitious force? What effect does it have on the atmospheric circulation?

An **inertial frame** is a reference frame in which Newton's law of inertia holds. Two different inertial frames move relative to each other uniformly.

We will describe most of the experiments in this book in the context of an inertial frame and will often refer to an **inertial observer**. This observer describes the experiment from an inertial frame.

Relative and absolute. The adjective *relative* in special relativity always means that a quantity depends on the reference frame (usually an inertial frame). The statement "the object is at rest" is relative. On the other hand, *absolute* means that a quantity is independent of the reference frame. In general, one restricts oneself to inertial frames. The statement "the charge is an absolute quantity" means that the charge of an object has the same value in all inertial frames.

[6] The equivalence relation "two reference frames move uniformly relative to each other" causes equivalence classes that partition the set of all reference frames. One of these classes is the inertial frames.

3.4 The Galilean Principle of Relativity

3.4.1 Introduction

For sure, in a train on a straight track, you will have already observed that, with closed eyes and covered ears, you hardly notice that the train moves. And most likely, you considered this to be trivial or self-evident. If this is the case, you have been pretty wrong. This observation expresses an important symmetry of Nature that restricts the set of possible laws of Nature considerably.[7] If one takes this observation really seriously, it almost leads to special relativity all by itself.[8]

We start with a **Gedanken experiment** and imagine that Alice and Bob are both in an inertial frame. Alice stands on a railway embankment and Bob sits in a train that moves uniformly with velocity v relative to Alice.

Alice, standing on the railway embankment, conducts experiments and investigates the free fall and the trajectories of thrown objects. She also conducts experiments in which gases are compressed, as well as other experiments using spring scales, and realizes that the results of all these experiments can be described with classical mechanics—without needing any fictitious force.[9]

Bob, sitting inside the moving train, conducts the same experiments as Alice. He also finds that classical mechanics is suitable for describing the results.

This important observation is formulated as a principle[10]:

> **Galilean principle of relativity (GPR)**: Mechanical processes in all inertial frames happen in the same way.

"Happen in the same way" means that they can be described with the same laws of classical mechanics (without fictitious forces). If Galilei's principle was wrong, Alice's experiments would behave according to classical mechanics, but Bob's experiments would involve fictitious forces or require completely different laws. Galilei's principle also means that, with mechanical experiments, two different inertial frames cannot be distinguished.

This is not true if one of the observers is not in an inertial frame. If Bob sits on the seat in the chairoplane, a relinquished object does not fall vertically down to the Earth. To describe the observed trajectory, he can still use classical mechanics but

[7] By *laws of Nature*, we refer to physical laws that describe Nature.

[8] The only missing piece is the fact that there is an absolute and *finite* velocity. This is the speed of light.

[9] The last statement is actually trivial, because the key property of inertial frames is that there are no fictitious forces.

[10] This principle was described first by Galileo Galilei in his epoch-making book *Dialogue Concerning the Two Chief World Systems*.

needs a fictitious force, the centrifugal force. This is how Bob notices that he is not in an inertial frame.

Focus again on Bob in the train. We stated that, by means of mechanical experiments, he cannot determine that he is in a uniformly moving train. Is this really true? He just has to look out of the window to see that he sits in a moving train! Isn't it easy to distinguish inertial frames? No. Remember that the idea is that Bob can describe his experimental finding using classical mechanics and without the need for fictitious forces. As we said, this is true in the case at hand, and the radiation from outside (which conveys the image of the outside of the train) is too small to influence the mechanical experiments. Alternatively, we can imagine that Alice and Bob conduct their experiments in an opaque box (but still in the railway embankment and train reference frames, respectively).

Because two inertial frames are equivalent, we can no longer tell if two events happen at the same location. The ticks of the clock located next to Bob in the moving train happen at the same location for Bob, but not for Alice, who stands on the railway embankment.

Exercise 9: Alice and Bob are in two different trains, but each is at rest in some inertial frame. How do these trains move relative to each other? Claire is in a third train that moves accelerated relative to the other trains. If Alice lets loose an object, in her reference frame, it falls vertically down to Earth. The same holds for Bob if he lets loose an object and describes its motion in his inertial frame. What happens, if Claire relinquishes an object in her train? In her reference frame? In Alice's or Bob's inertial frame?

Exercise 10: The velocity of an object always has to be given relative to a reference frame (or relative to another object, which defines a reference frame). In the case of acceleration, the reference frame often is omitted. Why is this possible and relative to which reference frame is the acceleration then meant to be understood?

3.4.2 Quantitative Description and Galilei Transformation

So far, our discussion of the experiments carried out by Alice and Bob has been purely qualitative. Now, we lay eyes on the **quantitative description**.

Coordinate systems. In order to do this, Alice and Bob must introduce coordinates (see Fig. 3.2). Imagine that Alice has a coordinate system with the coordinates x, z whose x-axis coincides with Bob's direction of motion (i. e., in the direction of the railway tracks) and her z-axis points vertically upward. Bob has a coordinate system with coordinates x', z'. His x'-axis coincides with Alice's x-axis and his z'-axis also points vertically upward. The origins of the coordinate systems are at Alice or Bob, respectively.

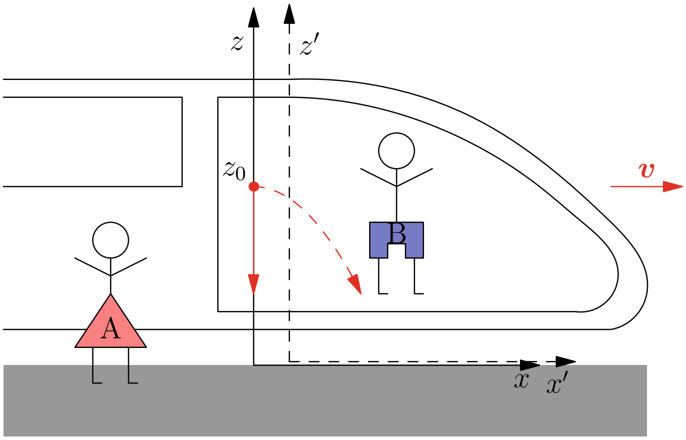


Fig. 3.2 Free fall in the train

Motion from the point of view of Alice and Bob. Now, Bob carries out a free-fall experiment. He describes the trajectory by

$$x'(t) = 0 \quad \text{and} \quad z'(t) = z_0 - \frac{1}{2}gt^2 \tag{3.1}$$

(with the time t that Bob, for instance, reads from the railway station clock). For Alice, Bob's free fall looks as if Bob had thrown the objects horizontally. She describes Bob's experiment with the trajectory

$$x(t) = vt \quad \text{and} \quad z(t) = z_0 - \frac{1}{2}gt^2\,. \tag{3.2}$$

The trajectories obviously are different. For Bob, it is a vertical (straight) line, and for Alice, half of a parabola. But there is only one object and one trajectory from different points of view, thus there must be a formula that calculates Alice's form of the trajectory from Bob's trajectory, or vice versa. The question is the following: suppose you have Alice's coordinates of a point P. How can we calculate Bob's coordinates of the same point? It is obvious that $x' = x - vt$ and $z' = z$ is the wanted formula.

And what about the time? In classical mechanics, the time is the same for both, Alice and Bob. There is only one time. This is called Newton's **absolute time**. But we can introduce a separate time t' for Bob for later use. Then, however, we always have $t' = t$. Just as Alice and Bob have different space coordinates, we also use different time coordinates.

Galilei transformation. Calculating the coordinates of an arbitrary point P in one coordinate system from the given coordinates in another coordinate system is called a **coordinate transformation**. The coordinate system under discussion here is the Galilei transformation:

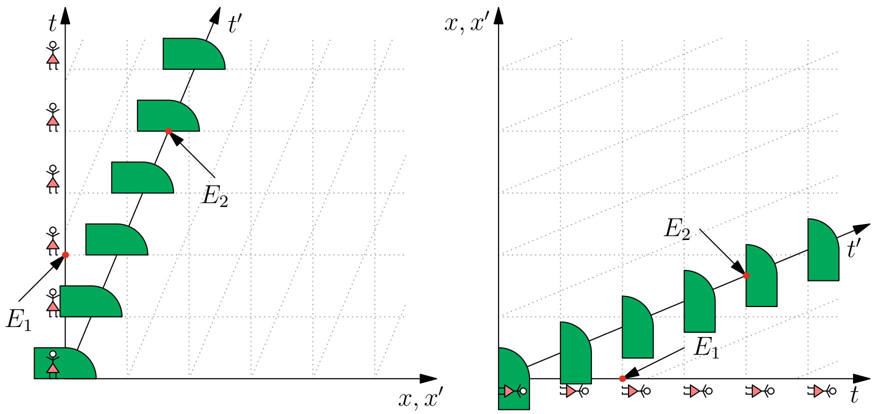


Fig. 3.3 Alice, Bob and the Galilei transformation. Left: t drawn over x; Right: x drawn over t

Galilei transformation (GT): Given the trajectory of an object with coordinates x, y, z and time t in coordinate system A, the coordinates x', y', z' and the time t' in coordinate system B are then given by

$$\begin{aligned} x' &= x - vt \\ y' &= y\ , \quad z' = z \\ t' &= t\ . \end{aligned} \tag{3.3}$$

Here, coordinate system B is equally oriented as coordinate system A and moves with velocity v in the x-direction relative to A.

Figure 3.3 clarifies the circumstances. On the left side, the time t is drawn as a function of the space coordinate x. Alice stands at $x = 0$ and Bob's train moves according to $x = vt$ away from Alice. Bob's position in his coordinate system is $x' = 0$, and $x' = x - vt$ holds. Because there's only one absolute time, we have $t' = t$.

We call a point in the x-t diagram an *event*. An event occurs at a fixed location in space and a fixed time.

With the Galilei transformation, Alice's coordinates of an event are transformed into Bob's coordinates, or vice versa. Consider the event E_1 in which Alice claps her hands. It is $E_1 : (t_1, x_1 = 0)$. According to the Galilei transformation (3.3), for Bob, it is $E_1 : (t_1', x_1' = x_1 - vt_1 = -vt_1')$. For the event E_2, in which Bob blinks his eyes, we have $E_1 : (t_2', x_2' = 0)$. Again, with the Galilei transformation, but this time solved for (t, x, y, z), we get $E_2 : (t_2, x_2 = x_2' + vt_2' = vt_2)$.

A further remark about Fig. 3.3. In the literature one special relativity, on mostly finds diagrams, in which the time is drawn over the position (as on the left side of the figure). But we will follow the practice in classical mechanics, in which the position is drawn over the time (right side of the figure).

So far, we have described the trajectory of a object from the point of view of Alice and that of Bob. Now, if the trajectory for one of them follows from Newton's laws, it must also hold for the other one. Otherwise, the inertial frames would be distinguishable and Galilei's principle of relativity would be violated.

Form-invariance of Newton's laws. Now, we derive the trajectory from both observer's points of view and show that **Newton's force law**

$$\begin{aligned} F_x &= m\ddot{x}(t) \\ F_z &= m\ddot{z}(t) \end{aligned}$$

is fulfilled in both cases.

For Bob, from (3.1), it follows that $\dot{x}'(t) = 0$ and $\ddot{x}'(t) = 0$, which means that, according to the law of motion, there must be no force in the x-direction on the falling object. This is consistent with our expectation, because the only acting force is the gravitational force, which has no component in the x-direction. Furthermore, $\dot{z}'(t) = -gt$ and $\ddot{z}'(t) = -g$. According to the force law, a force with magnitude $m'g$ must act in the vertical direction downward (m' is the mass of the falling object for Bob). This is exactly the gravitational force.

From Alice's point of view, the task is even easier. For her, from (3.2) it also follows that $\dot{x}(t) = 0$ and $\ddot{x}(t) = 0$, which again means that, according to the force law, there is no force in the x-direction. Furthermore, $\dot{z}(t) = -gt$ and $\ddot{z}(t) = -g$. According to the force law, a force of magnitude mg acts vertically downward (m is the mass of the falling object for Alice). This is again the same gravitational force. Now, we assume that the mass of the object is the same for Alice and Bob, i. e., $m = m'$. Then, for both observers, the magnitude of the gravitational force must be the same: $F_z = F'_{z'} = F_g$.

Therefore, in this particular case, Newton's force law has the same form for Alice and for Bob. This property is called **form-invariance**.[11] An equation is called form-invariant if it keeps its form when subjected to a transformation, here, a Galilei transformation.

Initial condition. For both Alice and Bob, the trajectory of the object fulfills Newton's force law. Still, the trajectories are different: for Alice, the object was thrown horizontally, and for Bob, it is in a free fall. How can that be? The secret is the **initial condition**. This is different for Alice and Bob: the initial position of the object is the same for Alice and Bob, but the initial velocity is different. For Bob, the object is at rest initially, and for Alice, it moves horizontally with $\dot{x}(0) = v$. And Newton's force law yields different trajectories for different initial conditions. But it is not possible to distinguish different inertial frames because Alice could equally

[11] *Invariance* may sound complicated, but is not. *Variance*, from the Latin *variantia*, means that something changes. The prefix *in* produces the contrary. An object is **invariant** with respect to a transformation, if it does not change with the transformation. A circle is invariant regarding a rotation by an arbitrary angle around the center of the circle. And the function $z = xy$ is invariant regarding exchanging the coordinates x and y.

conduct the free-fall experiment, which, for Bob, would look like a horizontally thrown object. In all experiments, Alice and Bob can be interchanged; this is required by Galilei's principle of relativity.

Conclusion. What have we achieved so far? For a special trajectory, we have shown that *if* the trajectory in Alice's experiment fulfills Newton's law of motion *and* Bob's trajectory can be calculated from Alice's trajectory by a Galilei transformation, *then* Bob's trajectory also fulfills Newton's force law.

It is also easily possible to show generally that *Newton's force law is form-invariant upon a Galilei transformation*. To do that, we start with Bob's "version" of Newton's force law, i. e., $F'_{x'} = m\ddot{x}'(t')$, and carry out a Galilei transformation. Forces and masses stay the same, so $F'_{x'} = F_x$ and $m' = m$, and therefore $F_x = m\ddot{x}'(t')$. Only the second derivation of the x'-coordinate of the trajectory with respect to the time is left. First, we have $t' = t$, then we must differentiate $x'(t) = x(t) - vt$ twice. From that, it follows that $\ddot{x}'(t) = \ddot{x}(t)$, and thus $F_x = m\ddot{x}(t)$ and we have reached our goal.

If an arbitrary trajectory of an object's motion fulfills Newton's force law $F_x = m\ddot{x}(t)$ for the inertial observer Alice, the trajectory of the same motion also fulfills Newton's law of motion $F_{x'} = m\ddot{x}'(t')$ for another inertial observer Bob. The coordinates of the trajectories are related by the Galilei transformation.

One says: **Newton's force law is form-invariant regarding the Galilei transformation**[12] (it does not change its form if one subjects it to a Galilei transformation).

Newton's first law (law of inertia) in this context can be considered a special case of Newton's third law. And due to the fact that forces do not change upon a Galilei transformation, Newton's third law (action-reaction law) is also form-invariant regarding the Galilei transformation. Therefore, all Newton's laws are form-invariant regarding the Galilei transformation.

Exercise 11: From astronomy, you know that the description of the motion of the planets is much easier in the heliocentric reference frame than in the geocentric one. This was determined by Nicolaus Kopernikus. In the heliocentric reference frame, the planets, according to Johannes Kepler, move along elliptic orbits, with the Sun in one of the focal points of the ellipse. For most planets (including the Earth), the ellipse almost has the form of a circle. Which form does a planet's trajectory have when described in the geocentric reference frame?

Exercise 12: Alice stands on the Earth's surface. As she does so, Bob and Claire both jump out of an airplane with parachutes at the same time. Describe the trajectory of Claire, once from Alice's reference frame and once from Bob's reference frame. Neglect the air resistance.

[12] We consider the transformation rules $\boldsymbol{F}' = \boldsymbol{F}$ and $m' = m$ as part of the Galilei transformation.

3.4.3 Summary

We started with the observation that mechanical processes happen in the same way in all inertial frames. Otherwise, one would be able to distinguish two inertial frames with mechanical experiments; they would not be equivalent. This means that Newton's laws must "look equal" for all inertial frames. In particular, there are no fictitious forces in any of the inertial frames.

Then, we determined the following: if Newton's force law holds for Alice, then it holds automatically for Bob, *provided* that the transformation of the coordinates (or of the physical quantities in general) from Alice's to Bob's reference frame is performed with the Galilei transformation. One can say that the equations of classical mechanics are **form-invariant** regarding the Galilei transformation. This means simply that when transforming the equations of classical mechanics from Alice's to Bob's coordinate system with the Galilei transformation, they do not change their form. The equation $F = m\ddot{x}(t)$ transforms to $F' = m'\ddot{x}'(t')$, and not to $F' = m'\ddot{x}' + q\dot{x}' + 5$ or something else. The laws of classical mechanics and the Galilei transformation are tightly related.

From Galilei's principle of relativity (GPR) and the form of the laws of classical mechanics follows the Galilei transformation (GT). If the Galilei transformation would not hold, either Galilei's principle of relativity or the equations of classical mechanics would be wrong. As a boolean formula:

$$\text{Classical mechanics AND GPR} \Rightarrow \text{GT}. \tag{3.4}$$

3.5 Addition of Velocities

One surprisingly fruitful topic is the addition of velocities. To dive in, we imagine a railway embankment with two parallel tracks (see Fig. 3.4). Alice stands on the embankment. A (slow) regional train, in which Bob sits, passes by. At the same time, a (fast) long-distance train passes by, in which Claire sits. Now, Alice measures the velocity of the regional train and gets $\boldsymbol{v}_{\mathrm{BA}}$. The first index (here: B) denotes the object whose velocity is measured and the second index (here: A) denotes the reference frame, in which the velocity measurement is made. Alice also measures the velocity of the long-distance train and gets $\boldsymbol{v}_{\mathrm{CA}}$. Not only does Alice perform velocity measurements, but Bob does as well. He gets $\boldsymbol{v}_{\mathrm{CB}}$ for the velocity of the long-distance train. We also know that, for Alice, the velocity $\boldsymbol{v}_{\mathrm{CA}}$ of the long-distance train is given by the velocity $\boldsymbol{v}_{\mathrm{BA}}$ of the regional train plus the velocity $\boldsymbol{v}_{\mathrm{CB}}$ of the

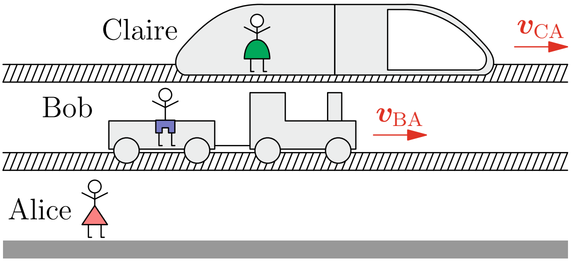

Claire
$\boldsymbol{v}_{\mathrm{CA}}$
Bob
$\boldsymbol{v}_{\mathrm{BA}}$
Alice

Fig. 3.4 Example for the addition of small velocities

long-distance train, as measured by Bob[13]:

$$\boldsymbol{v}_{\mathrm{CA}} = \boldsymbol{v}_{\mathrm{CB}} + \boldsymbol{v}_{\mathrm{BA}} \ . \tag{3.5}$$

This type of addition of velocities is called the **Galilean addition of velocities** (GAV). It is a vector addition. And the reason for the name is that it results from the Galilei transformation.

One thing is very important: here, the velocities that are added are *measured in different reference frames* ($\boldsymbol{v}_{\mathrm{BA}}$ was measured by Alice and $\boldsymbol{v}_{\mathrm{CB}}$ by Bob)! And this is why the statement (3.5) is not as trivial as it seems on first sight.[14]

From the formula (3.5), one recognizes that the choice of the notation is useful: on the right side of the equation, we have the indices (CB)(BA), i. e., equal indices “meet and cancel” and the indices (CA) of the left side of the equation are left over. You see that a clever choice of the notations helps us to remember equations and to uncover errors in calculations.

A *special case* of (3.5) is given when C = A. Then, it reads as $\boldsymbol{v}_{\mathrm{AA}} = \boldsymbol{v}_{\mathrm{AB}} + \boldsymbol{v}_{\mathrm{BA}}$. But because $\boldsymbol{v}_{\mathrm{AA}} = 0$, we conclude that $\boldsymbol{v}_{\mathrm{AB}} = -\boldsymbol{v}_{\mathrm{BA}}$. This means that Bob’s velocity, as measured by Alice, is equal up to a sign to Alice’s velocity, as measured by Bob.

You may think: ok, fine. But there’s nothing surprising about Equation (3.5). Why this fuss? As stated, the equation is not trivial. To the contrary: (3.5) is no longer valid for large velocities! We will see later why this is the case. To prepare, we derive the addition formula from the Galilei transformation.

[13] Note that we have broken away from our usual habit of denoting the physical quantities of Alice, Bob and Claire without primes, with one prime and with two primes, respectively. This is done for the purpose of improving the readability here.

[14] Strictly speaking, this is about the following: you get from one to the other coordinate system with a coordinate transformation that depends on the relative velocity of both coordinate systems. Instead of transforming directly with the coordinate transformation associated with $\boldsymbol{v}_{\mathrm{CA}}$ from Alice’s to Claire’s coordinate system, one can first perform a coordinate transformation with velocity $\boldsymbol{v}_{\mathrm{BA}}$ to Bob’s coordinate system, and then with a coordinate transformation with velocity $\boldsymbol{v}_{\mathrm{CB}}$ from Bob’s to Claire’s coordinate system. The sequence of two coordinate transformations is equal to one combined coordinate transformation, and this defines an operation on $\boldsymbol{v}_{\mathrm{BA}}$ and $\boldsymbol{v}_{\mathrm{CB}}$ that yields $\boldsymbol{v}_{\mathrm{CA}}$. And this operation becomes a simple vector addition only in the case of the Galilean addition of velocities.

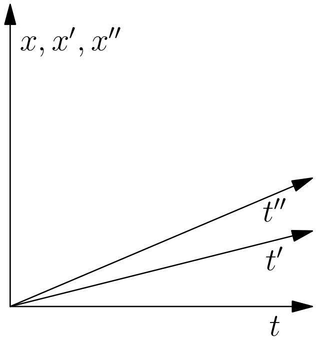


Fig. 3.5 To the derivation of the Galilean addition of velocities from the Galilei transformation

Derivation. Now, we come to the derivation of (3.5) from the Galilei transformation. We restrict ourselves to the case in which all velocity vectors lie on the same line and show that $v_{\mathrm{CB}} + v_{\mathrm{BA}} = v_{\mathrm{CA}}$. We start by giving coordinate systems to our observers: Alice gets (t, x), Bob gets (t', x'), and Claire gets (t'', x'') (see Fig. 3.5). The three x-axes coincide and, at the time $t = t' = t'' = 0$, all three observers meet in $x = x' = x'' = 0$. In addition, we assume that the three velocities in (3.5) all point in the positive x-direction. Thus, the Galilei transformation to transform Alice's to Bob's coordinates yields

$$x' = x - v_{\mathrm{BA}} t \ , \quad t' = t \ . \tag{3.6}$$

The location of the regional train is given by $x' = 0$, i. e., $x = v_{\mathrm{BA}} t$, for Alice. The Galilei transformation from Bob's to Claire's coordinates is

$$x'' = x' - v_{\mathrm{CB}} t' \ , \quad t'' = t' \ . \tag{3.7}$$

For the long-distance train, we have $x'' = 0$, i. e., $x' = v_{\mathrm{CB}} t'$, for Bob.

Finally, the Galilei transformation to transform Alice's to Claire's coordinates is

$$x'' = x - v_{\mathrm{CA}} t \ , \quad t'' = t \ . \tag{3.8}$$

And, as already stated, the location of the long-distance train is $x'' = 0$, i. e., $x = v_{\mathrm{CA}} t$, for Alice.

Equation (3.8) can be derived from Equations (3.6) and (3.7). To do so, we just put (3.7) into (3.6) and get

$$x'' = x' - v_{\mathrm{CB}} t' = (x - v_{\mathrm{BA}} t) - v_{\mathrm{CB}} t = x - (v_{\mathrm{BA}} + v_{\mathrm{CB}}) t \ , \quad t'' = t \ .$$

This must be equal to (3.8), which is exactly the case if $v_{\mathrm{CA}} = v_{\mathrm{CB}} + v_{\mathrm{BA}}$ holds. The formula (3.5) therefore is a consequence of the Galilei transformation, which brings us to an importance conclusion:

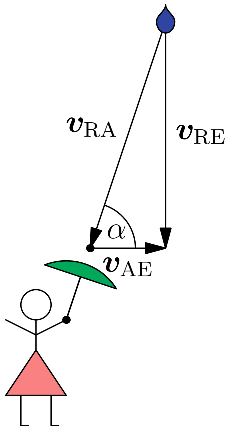

$\boldsymbol{v}_{\mathrm{RA}}$
$\boldsymbol{v}_{\mathrm{RE}}$
α
$\boldsymbol{v}_{\mathrm{AE}}$

Fig. 3.6 A falling raindrop, from the perspective of running Alice

The Galilean addition of velocities (GAV) follows from the Galilei transformation. As a boolean formula,

$$\mathrm{GT} \Rightarrow \mathrm{GAV} . \tag{3.9}$$

If the Galilean addition (3.5) of velocities (e. g., for large velocities) is violated, then the Galilei transformation cannot be valid.

But more about that later.

Example: running in the rain. Imagine that Alice is standing in the rain with her umbrella open. To protect herself from the rain, the umbrella is directly above her, as the rain comes straight down (there is no wind). Now, she starts running. As a result, she has to tilt her umbrella in order to stay dry.

Let us take a closer look at this effect (see Fig. 3.6). In the rest frame of the Earth's surface (E), the rain drops (R) fall exactly vertically, with a velocity of $v_{\mathrm{RE}} = 30\,\mathrm{km/h}$, which is a typical velocity for falling raindrops. Let us assume that Alice (A) is running with a velocity of $v_{\mathrm{AE}} = 6\,\mathrm{km/h}$. Then, the velocity of the raindrop in the reference frame where she is at rest is $\boldsymbol{v}_{\mathrm{RA}} = \boldsymbol{v}_{\mathrm{RE}} - \boldsymbol{v}_{\mathrm{AE}}$, which follows from $\boldsymbol{v}_{\mathrm{RE}} = \boldsymbol{v}_{\mathrm{RA}} + \boldsymbol{v}_{\mathrm{AE}}$.

Because of $\tan\alpha = v_{\mathrm{RE}}/v_{\mathrm{AE}} = 5$, we have $\alpha \approx 78.7°$. The raindrop, from Alice's point of view, comes from a direction of 78.7° above the horizon. The apparent location of an object (or its apparent origin), here, the source of the raindrops, is shifted relative to its true location (or its true origin) in the direction of the observer's motion.

A similar effect exists for light, in which case it is called an *aberration* and this will not be the last time you hear about it in this book.

3.6 Summary

In this chapter, we learned that the description of Nature is always **relative to a reference frame**. There are particular reference frames; these are the **inertial frames** in which the law of inertia holds. Inertial frames are also special because, in inertial frames, the laws of classical mechanics take their most simple form (i. e., **without fictitious forces**). The form of these laws is the same in all inertial frames. This is the content of the **Galilean principle of relativity**. It also means that, with mechanical experiments, two inertial frames cannot be distinguished. The transformation of physical quantities in classical mechanics from one inertial frame to another one is performed with the **Galilei transformation**. An important consequence of the Galilei transformation is the **Galilean addition of velocities**. If the Galilei transformation should be wrong, then either Galilei's principle of relativity or the laws of classical mechanics are wrong. And if the Galileian addition of velocities should be wrong, necessarily, the Galilei transformation is wrong.

Chapter 4
Waves and Light

4.1 Introduction

Light plays a central role in special relativity. The reason for this will become clear in the course of this book. According to electrodynamics, light is an electromagnetic wave. Therefore, we dedicate ourselves to waves in this chapter.[1]

In the second part of the 17th century, two different theories of light were developed. One was the **wave theory of light** (or wave optics) of the Dutchman Christian Huygens and the other was Isaac Newton's **corpuscular theory of light**. According to Huygens, light is a wave phenomenon, similar to a surface wave in (shallow) water or a sound wave in the atmosphere. And for Newton, light consisted of small light particles ("corpuscles").

It is not obvious that light should be a wave. Think about the beam that emanates from a laser pointer. This beam propagates along a (straight) line. A water wave, however, behaves completely differently. It propagates in all directions. Huygens was able to show, with his **principle of superposition of elementary waves**, that a light wave in the absence of obstacles may indeed propagate on (straight) lines. On the other side, Newton had a difficult time trying to explain light phenomena like light diffraction in a prism. Nevertheless, Huygens' theory was not taken seriously, and the reason for this was mainly Newton's authority. With his mechanics, he had acquired such an esteemed reputation that most of his contemporaries simply could not imagine that he could be wrong with his corpuscular theory of light.[2]

An important milestone on the path toward acceptance of the wave theory of light is dated to the year 1802. In that year, Thomas Young carried out his famous **double slit experiment** (you might possibly know it from quantum theory). The result was easily explained with the wave theory of light, something that the corpuscular theory

[1] Quantum theory is irrelevant for special relativity and will not be touched upon here.

[2] Possibly, you have heard that, according to quantum theory, light consists of photons, which share many properties with particles but others with waves. Photons do not behave like Newton's corpuscles. To say that Newton was right after all, would be quite absurd.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_4

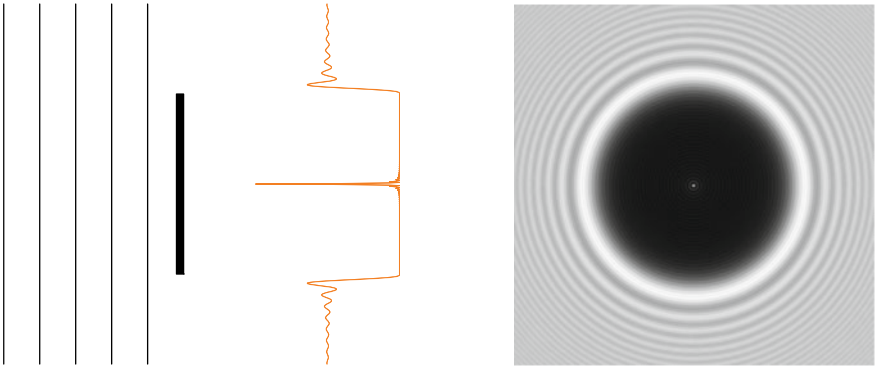


Fig. 4.1 Poisson spot. Left: A wave coming from the left hits a disk (black vertical bar). Far behind the disc, the indicated intensity distribution with the Poisson spot in the middle can be observed (orange). Right: Photography of the shadow behind the disc. The diffraction rings and the very small Poisson spot in the center can clearly be seen

was not able to do. Nevertheless, the critics of the wave theory still had the upper hand.

But it is an interesting story as to how that changed. In 1818, during a competition in Paris, Siméon Poisson made fun of the wave theory while studying a paper by Augustin Fresnel. He pointed out that, according to the wave theory, there should be a bright spot exactly in the center of the shadow behind a uniformly illuminated disc, which would be absurd [Lipson+]. François Arago, however, later actually demonstrated the **Poisson spot** (see Fig. 4.1). It results from the fact that the elementary waves emanating from the edge of the disk have the same path length to the center of the shadow, and therefore interfere constructively there. With this, the critics of the wave theory of light were finally silenced and wave optics established.

In the years 1861–1864, James Clerk Maxwell published the **basic equations of electrodynamics**, which were named after him.[3] These basic equations predict the existence of electromagnetic waves, which were experimentally discovered by the German Heinrich Hertz in 1886. As a special case, Maxwell's equations include wave optics and show that light is an electromagnetic wave.

We will dedicate ourselves in Sect. 4.2 to waves in classical mechanics and in Sect. 4.3 to light waves.

[3] Maxwell combined existing laws (Gauss's law, Faraday's law of induction, Ampére's circuital law), completed them with a new phenomenon, the displacement current, and formulated them with Faraday's idea of fields.

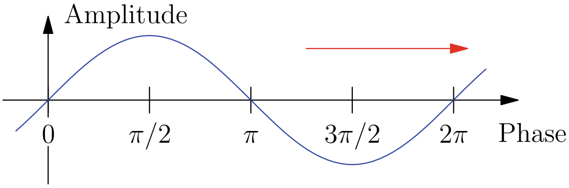


Fig. 4.2 A harmonic wave (sinusoidal wave) that moves in one dimension from left to right

4.2 Waves in Media

In the next section, Sect. 4.3, we will discuss light waves. To prepare for that, we start with a discussion of mechanical waves, because they are easier to grasp than light waves.

We start by discussing the main ideas about waves and then consider waves from different reference frames. Here, we encounter two fundamental effects, the **Doppler effect** and **aberration of waves**, which we subsequently discuss in more detail.

4.2.1 What Are Waves in Media?

Examples of waves in media. As examples of mechanical waves, we choose **surface waves in (shallow) water**. The effects that we observe in this system are transferable to other mechanical waves, e.g., **sound waves in the atmosphere**. The only important difference is that surface waves only propagate on a surface (in two dimensions) while sound waves clearly propagate in all directions in space (three dimensions).

Looking at a certain location in a wave, one observes that the **value of a certain physical quantity oscillates** in time. In the case of the surface wave in (shallow) water, the physical quantity is the height of the water compared to the average height, and in the case of sound waves, it is the pressure (or density) in comparison to the average pressure. In each space location in a wave, such oscillations occur. These oscillations at different locations are “coordinated” (or correlated) and form the wave in space.

The water and the atmosphere are the **medium** in which the wave phenomena occur. Initially, we assume that the **medium is at rest** in the inertial frame of the observer. In other words: stationary water and no wind.

Wave phenomena can be very complex. We consider only **linear** waves. This means that a sum (of the amplitudes) of two waves yields a new possible wave. The “sum” in the context of waves is also called **superposition**. For surface waves in (shallow) water and sound waves in the atmosphere, linearity is given if the amplitudes of the waves are not too large. Light waves in vacuum are strictly linear.

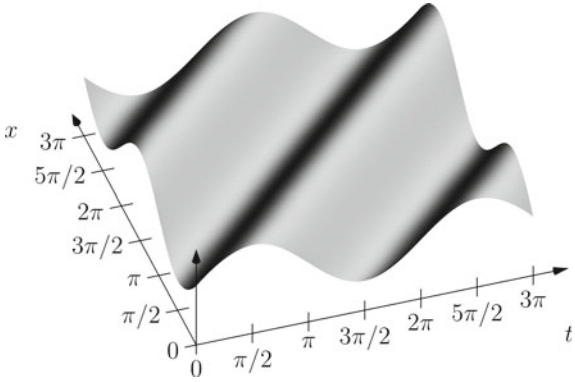


Fig. 4.3 A harmonic wave (sinusoidal wave) in the x-t-diagram. It travels in the positive x-direction

Waves in one dimension. Figure 4.2 shows a wave in **one dimension**, which propagates to the right at a certain moment.[4] It could be a water wave in a tight channel, a sound wave in a tight tube or even a wave on an infinitely long guitar string.

In the linear case, each such wave can be decomposed into **elementary waves** of the form

$$A(x,t) = A_0 \sin[2\pi \cdot (x/\lambda - \nu t) + \varphi_0], \tag{4.1}$$

that is an object which "lives" in the x-t-diagram, as shown in Fig. 4.3.

Here, A is the physical quantity that oscillates and A_0 the wave's amplitude. Further, ν is the **frequency** of the elementary wave, $T = 1/\nu$ its **period** and λ the **wavelength**. Often, the **angular frequency** $\omega = 2\pi\nu$ is used instead of the frequency and the **wave number** $k = 2\pi/\lambda$ instead of the wavelength. The frequency ν, the angular frequency ω and the period T are always positive.

The advantage of using ω and k instead of ν and λ is that we get rid of the factor 2π in the expression for $A(x,t)$ and can write

$$A(x,t) = A_0 \sin(kx - \omega t + \varphi_0). \tag{4.2}$$

If there are no possible misunderstandings, the angular frequency ω is often simply called "the frequency ω".[5]

Furthermore, φ_0 is a **phase shift**, which is important in the decomposition of waves into elementary waves. Here, it plays no role, thus we ignore it and set $\varphi_0 = 0$. The expression $\varphi = kx - \omega t$ is the **phase** of the wave (see Fig. 4.2). Locations of the

[4] We only consider propagating (traveling) waves. One example of a non-propagating wave would be if you were to pluck a guitar string. This is a standing wave.

[5] In order to avoid misunderstandings, the difference between the frequency and the angular frequency is usually made clear by a consistent choice of the symbols. The letters ν or f stand for the frequency and ω stands for the angular frequency.

wave with $\varphi = 0$ or $\varphi = \pi$ are called **wave nodes**. Furthermore, at $\varphi = \pi/2$, there is a **wave crest** and at $\varphi = 3\pi/2$ a **wave trough**.

As the decomposition of general (linear) waves into elementary waves is possible for all waves, elementary waves are often the only ones discussed.

Because of

$$\varphi = kx - \omega t = k\left(x - \frac{\omega}{k}t\right),$$

we see that the elementary waves depend on position and time only via the combination

$$\varphi = k\left(x - \frac{\omega}{k}t\right) = k\left(x - v_\mathrm{p}t\right),$$

the amplitude for fixed $x - v_\mathrm{p}t$ being constant. The elementary wave as a whole moves to the right with the **phase velocity** v_p. We have

$$v_\mathrm{p} = \frac{\omega}{k} = \lambda\nu. \tag{4.3}$$

The phase velocity v_p is the velocity of points with a fixed phase.

In the case of sound waves in the atmosphere, the phase velocity depends on the air pressure, and in the case of a guitar string, it depends on the thickness and the elasticity of the string. We assume that the air pressure is the same everywhere the wave moves[6] and that the guitar string has the same uniform thickness and elasticity. Then, one would say that **the medium is homogeneous**.[7] We will additionally call the phase velocity the **wave velocity** or the **propagation velocity of the wave** and refer to it with c_W (the c refers to the Latin “celeritas”, which means speed).

If the phase velocity is different for elementary waves of different frequencies, this is called **dispersion**. Then, the relation between ω and k, the **dispersion relation** $\omega(k)$, is no longer linear. In this case, waves that are composed of elementary waves of different frequencies do not retain their shape, but rather diffuse in space. We consider only waves that show no dispersion (dispersionless waves).[8]

The fact that we can communicate by talking shows that sound waves in the atmosphere and in the range of frequencies that we use for spoken communication are practically dispersionless. Otherwise, the sound waves that carry what is said to the listener would be deformed strongly on their way and the words could no longer be understood. Sound waves in the atmosphere can therefore be thought of as rigid waves of fixed shape moving at the speed of sound. Fig. 4.4a shows the shape of the sound wave corresponding to the spoken word “wave”.

[6] This is clearly not the case in the vertical direction in the atmosphere. For our purposes, however, this approximation is fine.

[7] “Homogeneous” comes from the Greek and means “unique” of “of uniform structure” (*homios* = same, *genos* = kind).

[8] While light waves in vacuum show no dispersion, they do so in matter. This is the reason why a prism splits a white light beam into its colored constituents and the sky is blue and the sunset red.

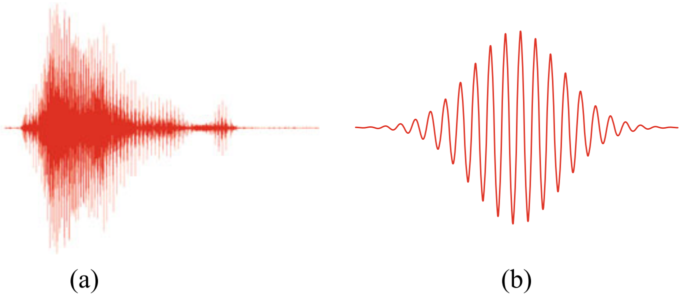
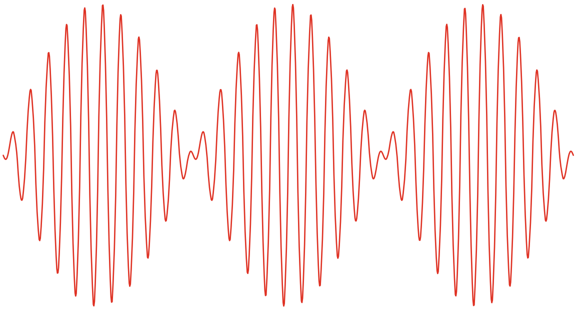


Fig. 4.4 **a** Left: A sound wave corresponding to the spoken word "wave"; **b** Right: A wave packet. The sound corresponds very roughly to a bang

Fig. 4.5 Two elementary waves added (or *superposed* or *interfered*) yields a beat pattern



A bang clearly is no elementary wave, but very roughly looks like the wave depicted in Fig 4.4b. This is a **wave packet**, which is composed of many elementary waves, all of them with a frequency close to a central frequency ν_0. A wave packet travels with the **group velocity** v_g, which, for dispersionless waves, is equal to the phase velocity v_p (the difference between the phase and the group velocity is somehow subtle and the statement holds only for the case of one dimension; see Sect. 4.5.3). In the case of sound waves in the atmosphere, both velocities are equal to the **speed of sound** c_S. In dry air at a temperature of 20 °C, the speed of sound is $c_S = 343$ m/s and practically does not depend on the air pressure. This velocity is relative to the medium (the air) at rest.

In case of dispersion, the group velocity becomes different from the phase velocity (usually smaller) and the wave packet becomes broader in time, it diffuses.

This can already be seen with a simple example. Take two elementary waves with the same amplitude and slightly different frequencies ω_1, ω_2, as well as wave numbers k_1, k_2, and add them. This yields

$$\sin(k_1 x - \omega_1 t) + \sin(k_2 x - \omega_2 t) = 2 \sin\left(\frac{\Delta k}{2} x - \frac{\Delta\omega}{2} t\right) \sin(\bar{k} x - \bar{\omega} t),$$

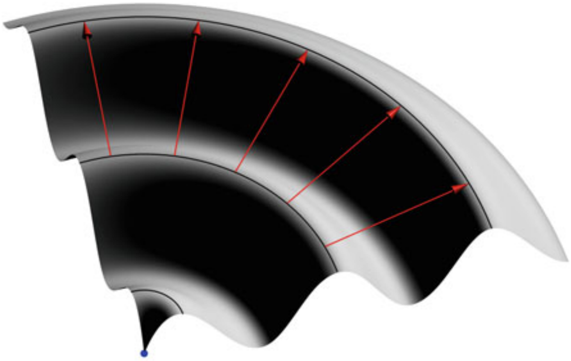


Fig. 4.6 Wavefronts and the propagation direction of a radially propagating wave

where $\Delta\omega = \omega_2 - \omega_1$, $\Delta k = k_2 - k_1$ are the differences of the frequencies and wave numbers, respectively, and $\bar{\omega} = (\omega_1 + \omega_2)/2$ and $\bar{k} = (k_1 + k_2)/2$ are the average frequency and wave number, respectively. The resulting wave at a certain moment looks like the example in Fig. 4.5 and is called a *beat pattern*. One recognizes a fast oscillation that corresponds to the factor $\sin(\bar{k}x - \bar{\omega}t)$ and describes that the wave nodes move with the phase velocity $v_\mathrm{p} = \bar{\omega}/\bar{k}$ and an "overall shape" or **envelope** of the wave that corresponds to the factor $\sin\left(\frac{\Delta k}{2}x - \frac{\Delta\omega}{2}t\right)$ and moves with the group velocity $v_\mathrm{g} = \Delta\omega/\Delta k$. If the dispersion relation is linear, the shape of this wave prevails in time; otherwise, it will smoothen out.

Waves in two or three dimensions. In **two or three dimensions**, in addition to the propagation velocity, the wave also has a **propagation direction**. Figure 4.6 shows a water wave. The black curves are lines of constant phase (wave crest, wave trough, wave node, or the like). Such curves are called **wavefronts**. In three dimensions, wavefronts are surfaces.

> The propagation direction of a wave (i.e., the phase velocity vector) is always perpendicular to the wavefronts.

The propagation velocity can depend not only on the location, but also on the direction in which the wave moves. If this is not the case, we speak of an **isotropic medium**.[9] The water that uniformly flows down a river is homogeneous for an observer resting on the riverbank, but not isotropic, because a wave traveling down the river is faster for this observer than a wave traveling up the river.

We assume that the medium of our waves is homogeneous and isotropic in the rest frame of its medium. Strictly speaking, for sound waves in the atmosphere, this is not the case (see Footnote 6). In the vertical direction, the air density (or pressure) changes according to the barometric formula. Directly above a dark surface (street

[9] "Isotropic" comes from the Greek and means "having the same properties in all directions" (*isos* = equal, *tropos* = direction, turn).

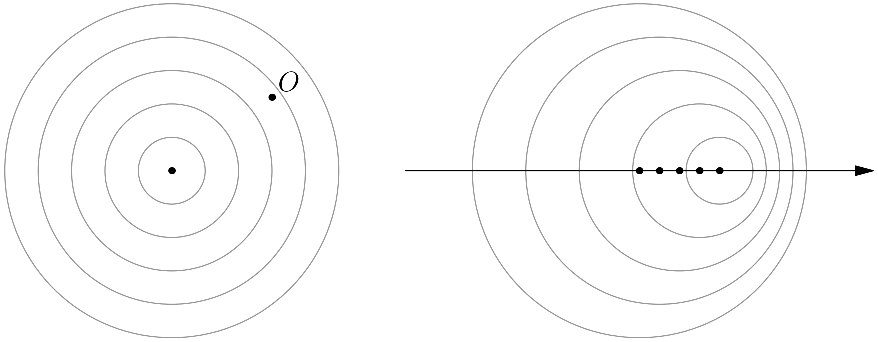


Fig. 4.7 A radially propagating wave. Left: Source is at rest relative to medium and observer. Right: Source moves relative to medium and observer

pavement) and for direct solar irradiation, these pressure differences are relatively large and, e.g., lead to a *fata morgana*.

The propagation velocity of the wave (magnitude and direction) as a vector is denoted by $\boldsymbol{c}_{\mathrm{W}}$. As mentioned, in a homogeneous and isotropic medium, it does not depend on the location or the direction.

For elementary waves in two or three dimensions, one usually chooses plane waves. **Plane waves** are very special waves, whose wavefronts are planes. Therefore, the propagation direction of plane waves is the same everywhere in the wave. For plane waves, the propagation velocity $\boldsymbol{c}_{\mathrm{W}}$ does not depend on the location. This statement is equivalent to that that wavefronts are lines or planes.

4.2.2 Waves in Moving Media: Qualitative Discussion I

Consider a medium that is homogeneous and isotropic in its rest frame and in which a wave propagates. We discuss three cases in which the source and the observer are at rest relative to the medium or the source or the observer move relative to it. All relative motions are uniform, accelerations play no role, and the scenery that we use is the inertial frame of the medium. We describe all phenomena from the point of view of the medium, which is at rest, by definition.

Source and observer are at rest. The easiest case is when both, the source and the observer are at rest. Under the assumption made above:

> The magnitude of the propagation velocity c_{W} **is independent of the location and the direction**.

How does this look for a water wave? Imagine Alice, who is at rest relative to the water, dipping a pen (the source) periodically into the water (see Fig. 4.7 on the left side). This creates circular wave crests and troughs (wavefronts), which spread

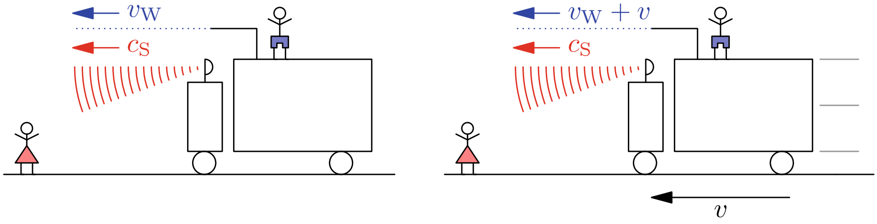


Fig. 4.8 Fire truck with a water jet pipe and a signal horn

concentrically around the source, and confirms the statement above regarding the propagation velocity.

Now, wherever the observer O sits, she measures a fixed frequency ν and a fixed wavelength $\lambda = c_W/\nu$ for the wave. The wavelength is the shortest distance between two wavefronts with $\varphi = 0$ and $\varphi = 2\pi$, respectively (or, more generally, with a phase difference of 2π). Furthermore, the velocity of the wave points away from the source.

Source moves, observer is at rest. The second case that we consider is when the source moves (relative to the medium). The observer is still at rest.

Velocity of the wave. Imagine that Alice walks with constant velocity along a straight bridge over the water and periodically dips her pen into the water. The wavefronts are still circular, but, as a family of circles, are no longer concentric (see Fig. 4.7 on the right side). The fact that they are circular and their center is the location where they have been created shows, that relative to the water's surface, the waves still propagate with the same velocity c_W in all directions. Therefore:

> The **velocity of the wave** (relative to the medium), which is created by a moving source, **does not depend on the velocity of the source**!

No matter how fast the source moves relative to the water's surface, the wave always moves with the velocity c_W relative to the that surface. *The medium determines the velocity of a wave, not the source.*

Wave velocity vs. particle velocity. Obviously, this statement is specific to waves in media. An example may help to clarify this. Consider a **fire truck** (see Fig. 4.8) with a water jet pipe and a signal horn. Initially, the truck is at rest relative to the observer Alice (and to the street). The signal horn is switched on and sends an alarm signal with a frequency of 440 Hz and a wave velocity equal to the speed of sound c_S. In addition, the water jet pipe is active and emits a water jet with a velocity of $v_W = 20\,\text{m/s} = 72\,\text{km/h}$ (this is the muzzle velocity relative to the street). Then, the fire truck moves slowly with a velocity of $v = 36\,\text{km/h}$ directly toward Alice, with the signal horn and the water jet pipe active.

What does Alice observe? First, the water jet, for her, has a velocity larger than v_W, because v_W is the velocity of the water jet relative to the water tube and, because of the Galilean addition of velocities, the velocity v of the fire truck has to be added to that of the water jet. Therefore, the water jet's velocity relative to Alice is $v'_\mathrm{W} = v_\mathrm{W} + v = 30\,\mathrm{m/s}$. For the velocity of the sound wave, however, nothing changes. The wave's velocity c_S is relative to the medium (the atmosphere):

> For **particles** (the water drops of the water jet), the velocity is always **relative to the source**, while for **waves**, it is **relative to the medium**.

Doppler effect. There is, however, another effect in the case of waves. If Alice measures the frequency of the signal as the fire truck moves with velocity v toward her, she gets 495 Hz, which is considerably higher than the 440 Hz emitted by the signal horn. This is the **Doppler effect**,[10] which is a consequence of the motion of the source (or, as we will see, the observer) relative to the medium.

In Fig. 4.7 on the right side, one recognizes the Doppler effect immediately: first, the distance between two wave crests (i.e., the wavelength λ) depends on the propagation direction of the wave and the velocity of the source. Now, the wavelength λ and the frequency ν of a wave depend, via $\nu\lambda = c_\mathrm{W}$, on the propagation velocity. Due to the fact that the propagation velocity c_W relative to the water is independent of the direction, the frequency ν of the wave (at a location fixed relative to the water) must depend on the velocity of the source and the propagation direction of the wave. We will return to the Doppler effect in Sect. 4.2.4.

Retardation. Another interesting effect for moving sources is that the direction of the wave's velocity (at the location of the observer) and the direction, in which the source is located, are different in general. This is shown in Fig. 4.9. At t_1, the observer O sees a wavefront and uses it to determine the wave's velocity $\boldsymbol{c}_\mathrm{W}$. This wavefront has been emitted by the source at time $t_0 = t_1 - T_{AO}$ when it was at location A. The traveling time T_{AO} and the distance d_{AO} between O and A is given by $d_{AO} = c_\mathrm{W} T_{AO}$. Since the creation of the wave crest, the source has moved, and at time t_1, it is at location B. *The source is not where you see it!* This effect is called **retardation**.

> The direction in which the observer sees the source (the *apparent direction*) and the *true direction* of the source in general are different.

Source is at rest, observer moves. Now, we consider the situation when the observer moves relative to the medium and the source is at rest.

Velocity of the wave. In Sect. 3.5, we have discussed the **aberration of raindrops**. It rains. In the rest frame of the Earth's surface,[11] the raindrops fall vertically with a velocity of $\boldsymbol{v}_\mathrm{RE}$. Alice, however, runs, and she has to tilt the umbrella in order to stay dry. If Alice's velocity (relative to the Earth's surface) is $\boldsymbol{v}_\mathrm{AE}$, the velocity of the raindrops in the reference frame of the running person is $\boldsymbol{v}_\mathrm{RA} = \boldsymbol{v}_\mathrm{RE} - \boldsymbol{v}_\mathrm{AE}$ (see Fig. 3.6). This follows from the Galilean addition of velocities. The direction from which the raindrops come shows aberration: *different observers see the raindrops coming from different directions*.

Consider a similar case with waves in media. A long swimming pool (see Fig. 4.10) is full of water at rest. At one end of the pool, Alice creates a wave that travels through

[10] Named after the Austrian physicist Christian Doppler (1803–1853).

[11] Or, to be more precise, to the atmosphere.

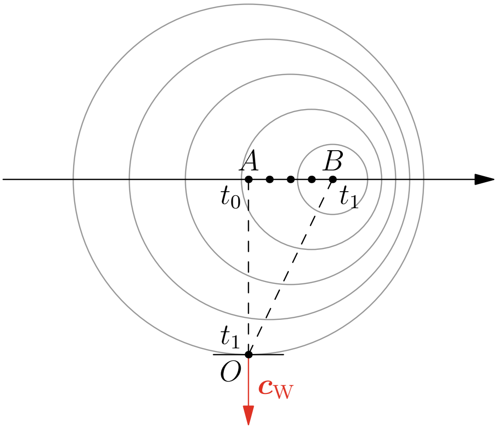


Fig. 4.9 Retardation. The direction AO from which the light comes and the direction BO of the source's location do not coincide

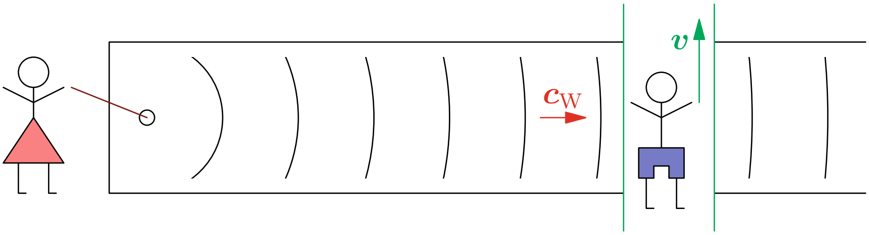


Fig. 4.10 To the aberration of waves in media

the pool with the velocity c_{W} for Alice. At some distance from Alice is Bob, who stands on a bridge that crosses the pool in a perpendicular manner. Bob determines the velocity $\boldsymbol{c}_{\mathrm{W}}{}'$ of the wave. He recognizes that the wavefronts are orthogonal to the pool, concludes that the propagation velocity points in the direction of the pool and deduces that $\boldsymbol{c}_{\mathrm{W}}{}' = \boldsymbol{c}_{\mathrm{W}}$. Now, Bob walks with velocity $\boldsymbol{v}$ perpendicular to $\boldsymbol{c}_{\mathrm{W}}$ along the bridge and again determines the propagation velocity of the water wave. The wavefront is parallel to his path, and therefore the propagation velocity of the wave is again perpendicular to his path and the bridge. The observer Bob at rest relative to the pool sees the wave coming from the same direction as the observer Bob who moves perpendicular to the pool. This is completely different from the case of the raindrops. *The propagation velocity of the wave shows no aberration.* The reason for this is that the wavefront is a geometric quantity:

Wavefronts (in classical physics) show no aberration.

For a vector to be a velocity, we demand that, on occasion of a reference frame change, it transfers according to the Galilean addition of velocities. The phase velocity does not do so. Therefore, in this sense, *the phase velocity is not a velocity.*

Suppose that Bob now walks with velocity $\boldsymbol{v}$ along the edge of the pool toward Alice. If he measures the wavelength, he gets the same result as Alice. But if he measures the period of the wave, he gets a value smaller than T. Therefore, the phase velocity $c_{\mathrm{W}}{}' = \lambda \nu = \lambda/T$ is larger for him. He gets $\boldsymbol{c}_{\mathrm{W}}{}' = \boldsymbol{c}_{\mathrm{W}} - \boldsymbol{v}$. In the direction perpendicular to the wave, the phase velocity indeed depends on whether he moves or not.

Doppler effect. This is straightforward now. In the case when the source moves and the observer is at rest (relative to the medium), we have seen that, for the case when the source moves towards the observer, the wavelength is smaller and the frequency higher for the observer.

Imagine now again the situation in Fig. 4.10 with Bob walking toward Alice. Bob travels towards the wavefronts, and therefore the wavefronts arrive more frequently at his location, which means that the frequency is again larger.

4.2.3 Waves in Moving Media: Qualitative Discussion II

Aberration and the Doppler effect. When discussing the person walking in the rain, we have first encountered the aberration phenomena. There is a source, the cloud, which sends particles (raindrops) toward an observer. We found that the direction from which the particles come depends on the velocity of the observer. We can describe this in a different way, with **two observers**. In the example of the raindrops, one observer would be at rest relative to the Earth's surface while the other would move with a velocity v relative to it. Then, the aberration phenomenon would be that the two observers see the raindrops coming from different directions. In the explanation given for the raindrops, the difference in directions (or **aberration angle**) would be a result of the Galilean addition of velocities.

Aberration is *very different* from the Doppler effect – and not only because aberration is about direction and the Doppler effect about frequency. In the case of the Doppler effect, the velocity of the source and the velocity of the observer both matter (both relative to the medium). The Doppler effect is an effect "between" the source and the observer (and of their velocities relative to the medium).

In the case of the aberration, the velocity of the source relative to the medium plays no role. Here, it is the difference between the velocities of the two observers (and their velocities relative to the medium) that matters. *Aberration is an effect "between" two observers.*

Aberration for wave packets. Unfortunately, we have seen that the phase velocity shows no aberration, or, to be more precise: it shows a different type of aberration than do particles (i.e., the raindrops). This is a problem, because we will see later (in Sect. 4.4), that the aberration of light from stars cannot be explained with the aberration effect of the phase velocity and that the aberration of raindrops works much better, although not perfectly, to explain the aberration of starlight. But light is a wave, and this means that we need a different explanation for the aberration of waves.

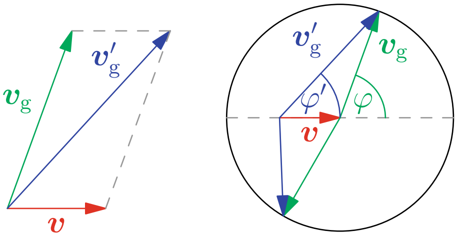

$\boldsymbol{v}_\mathrm{g}$
$\boldsymbol{v}'_\mathrm{g}$
$\boldsymbol{v}$
$\boldsymbol{v}'_\mathrm{g}$
$\boldsymbol{v}_\mathrm{g}$
φ'
φ
$\boldsymbol{v}$

Fig. 4.11 Wave packets in moving media. Left: Vector addition of group velocities; Right: Direction-dependency of the group velocity

The key is **wave packets**. We have seen that there are not only *extended* (or non-localized) waves (e.g., the elementary waves), but also *wave packets* that may be relatively well localized and almost resemble particles.

It turns out the the velocity of wave packets (i.e., the group velocity) transfers according to the Galilean addition of velocities if the inertial frame is changed. This is the same as for raindrops. There is one important difference, however. In the rest frame of the medium, the group velocity is independent of direction and location.

Suppose that Alice is at rest relative to the medium. Then, the velocity of a wave packet traveling in the direction $\boldsymbol{e}$ is given by $\boldsymbol{v}_\mathrm{g} = v_\mathrm{g}\boldsymbol{e}$, with a fixed magnitude v_g.

Now, let Bob move with velocity $-\boldsymbol{v}$ relative to the medium and Alice. From Bob's perspective, Alice and the medium move with velocity $\boldsymbol{v}$ relative to him. The wave packet that, for Alice, travels in the direction $\boldsymbol{e}$, for him, has the velocity

$$\boldsymbol{v}'_\mathrm{g} = \boldsymbol{v}_\mathrm{g} + \boldsymbol{v} = v_\mathrm{g}\boldsymbol{e} + \boldsymbol{v}, \tag{4.4}$$

which is visualized in Fig. 4.11 on the left side. In the figure on the right side, starting from the center of the circle, the propagation velocity v_g is drawn. It does not depend on the direction; the tip of the vector lies on a circle. The vector of the propagation velocity $\boldsymbol{v}'_\mathrm{g}$ for Alice has its tail at a point that is shifted by the vector $\boldsymbol{v}$ from the center of the circle. Its tip also lies on the circle, and therefore the magnitude (length) v'_g of this vector depends on its direction.

In the reference frame that moves (uniformly) relative to the medium, the magnitude of the group velocity depends on the direction.

Again, the magnitude of the propagation velocity of a wave is independent of its direction **only** in a **special inertial frame**,[12] which is the inertial frame in which the homogeneous and isotropic medium is at rest. A theory that states that the propagation velocity of the waves does not depend on the direction, can only be valid in a special inertial frame. This is because of the Galilean addition of velocities. The statement "the sound velocity (at normal conditions) is 343 m/s" can hold only in this special inertial frame, which is the rest frame of the medium. A "relativity principle" similar to that in classical mechanics cannot exist for such a theory.

From (4.4), we also see that the direction of the velocity of a wave packet from Alice's point of view is a different one than in the rest frame of the medium. This

[12] We presume that the medium is not accelerated.

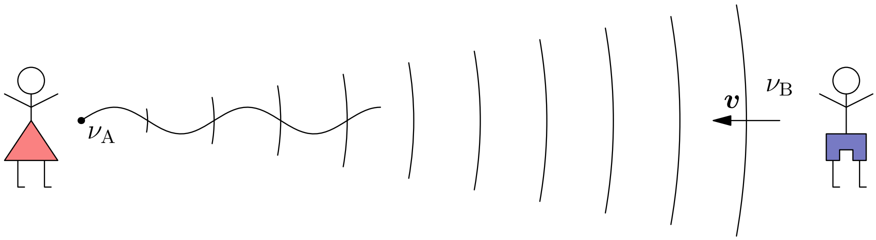


Fig. 4.12 Change of the frequency for a fixed source and a moving observer (relative to the medium). The curves are wavefronts

is the **aberration** of waves. The difference $\delta = \varphi' - \varphi$ of the angles φ' and φ (see Fig. 4.11) is the aberration angle.

We will see that, with *wave packets*, we can gain a *first understanding of the aberration of starlight*. But the situation with waves in media is unsatisfying. We have elementary waves that move with the phase velocity and wave packets that move with the group velocity. And these velocities transform in different ways: one does transform according to the Galilean addition of velocities, the other one does not. If both velocities point in the same direction in one reference frame, they won't necessarily do so in a different reference frame.

The special theory of relativity will resolve this problem, and we discuss this in detail in Sect. 12.5. Insofar, we will assume in this book that the propagation velocities for the considered waves show aberration (as wave packets do) and behave according to the Galilean addition of velocities. Therefore, (4.4) holds, and *we will now understand the wave velocity $\mathbf{c}_W$ to be the group velocity of a wave packet.*

4.2.4 Doppler Effect

Overview. In Sect. 4.2.2, we have considered the situation when the observer is at rest relative to the wave's medium but the source moves. The observer then measures a frequency different from that created by the source (which also depends on his position relative to the source and its motion), and we have encountered the same effect when the source is at rest and the observer moves. We now investigate this *Doppler effect* more systematically using the example of sound waves that propagate in the resting atmosphere with the speed of sound c_S.

As you would expect, our actors Alice and Bob will be with us. Alice drives the source and Bob is the observer.

Moving observer. We start with the case of a *moving observer*. In Fig. 4.12, you see Alice with the source A, which is at rest relative to the medium. Furthermore, there is Bob, the observer B. The source produces a sound wave with the frequency ν_A. Once in the time interval $\Delta t_A = 1/\nu_A$ (the wave's period), the amplitude crosses zero from negative to positive values. This is indicated by the wavefronts in the figure.

The question is which frequency ν_B the observer Bob measures. If he does not move relative to the medium, the wavefronts arrive at his location with period $\Delta t_B = \Delta t_A$ and therefore $\nu_B = \nu_A$.

If Bob moves toward the source, *he runs into the wavefronts*. The period Δt_B between two subsequent wavefronts then is smaller than Δt_A, and the frequency ν_B measured by the observer is therefore larger than ν_A. If Bob moves away from the source (this means $v_A < 0$), the frequency ν_B is smaller than ν_A. We will demonstrate in a minute that

$$\nu_B = \nu_A \cdot \frac{c_S - v_B}{c_S}, \tag{4.5}$$

where v_B is the velocity of the observer Bob B relative to the medium (the sign of v_B is negative if the observer B moves toward the source A and positive in the other case) and c_S the speed of sound.[13]

Moving source. Instead of having the observer moving relative to the source, we can also *move the source and have the observer be at rest* (both relative to the medium). Figure 4.7, right side, shows what happens then. If the source at a particular location (in the reference frame of the medium) creates a wave, then the wavefronts propagate radially. But due to the fact that the source moves, the next wavefront starts from a different location than the former. The wavefronts are not concentric anymore. If the source moves toward the observer Bob, the latter measures a frequency ν_B that is larger than ν_A. And if the source moves away from the observer B, the frequency ν_B is smaller than ν_A. We have

$$\nu_B = \nu_A \cdot \frac{c_S}{c_S - v_A} \tag{4.6}$$

(the sign of v_A is positive if the source A moves toward the observer B and negative in the other case).

Derivation and the general case. It is time now to derive the formula for the general case, in which both, the source Alice and the observer Bob move relative to the medium. We describe the situation in the rest frame of the medium (see Fig. 4.13). Alice and Bob move away from each other in opposite directions. The trajectory of the observer Bob is given by

$$x_B(t) = x_{B,0} + v_B t,$$

and that of Alice with the source is

$$x_A(t) = x_{A,0} + v_A t.$$

Here, we assume that $x_{B,0} > x_{A,0}$. In the figure, we have additionally assumed that $v_A < 0 < v_B$, which is not necessary for the derivation.

[13] We suppose that $|v_B| < c_S$. For $v_B = c_S$, Eq. (4.5) is still correct. $\nu_B = 0$ then means that the wavefronts for Bob do no longer move. For $v_B > c_S$, we get $\nu_B < 0$. The frequency in this case, however, is still larger than zero, but the traveling direction of the wave changes.

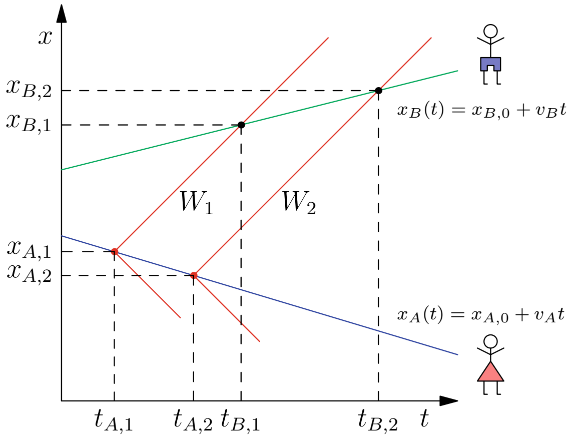


Fig. 4.13 For the derivation of the Doppler effect. The blue line represents the source and the green line the observer. The red lines show the motion of the wavefronts of the wave produced by the source

The source emits a wave. At time $t_{A,1}$, the wavefront W_1 with phase $\varphi = 0$ leaves the source. The next wavefront W_2 with $\varphi = 2\pi$ leaves the source at time $t_{A,2}$. For the source, the frequency of the wave is given by

$$\nu_A = \frac{1}{t_{A,2} - t_{A,1}}.$$

The wavefronts move toward the observer Bob. They arrive there at $t_{B,1}$ and $t_{B,2}$, respectively. Bob measures the frequency

$$\nu_B = \frac{1}{t_{B,2} - t_{B,1}}.$$

To determine this frequency, we have to express the times $t_{B,1}$ and $t_{B,2}$ of the observer as a function of the times $t_{A,1}$ and $t_{A,2}$ of the source.

The wavefront W_1 has the trajectory

$$\begin{aligned} x(t) &= c_\mathrm{S} \cdot (t - t_{A,1}) + x_{A,1} \\ &= c_\mathrm{S} \cdot (t - t_{A,1}) + x_{A,0} + v_A t_{A,1} \\ &= c_\mathrm{S} t + x_{A,0} + (v_A - c_\mathrm{S}) t_{A,1}. \end{aligned}$$

We determine the t-coordinate $t_{B,1}$ of the intersection point of this trajectory with that of Bob:

$$\begin{aligned} & x_{B,0} + v_B t_{B,1} = c_\mathrm{S} t_{B,1} + x_{A,0} + (v_A - c_\mathrm{S}) t_{A,1} \\ \Longrightarrow\ & x_{B,0} + (v_B - c_\mathrm{S}) t_{B,1} = x_{A,0} + (v_A - c_\mathrm{S}) t_{A,1}. \end{aligned} \tag{4.7}$$

We perform the same for the wavefront W_2 and get

$$x_{B,0} + (v_B - c_S)t_{B,2} = x_{A,0} + (v_A - c_S)t_{A,2}. \tag{4.8}$$

Now, we only have to subtract (4.8) from (4.7), which yields

$$(c_S - v_B)(t_{B,2} - t_{B,1}) = (c_S - v_A)(t_{A,2} - t_{A,1}).$$

For the ratio of the frequencies $\nu_B/\nu_A = (t_{A,2} - t_{A,1})/(t_{B,2} - t_{B,1})$, we immediately get

$$\frac{\nu_B}{\nu_A} = \frac{c_S - v_B}{c_S - v_A} \qquad (\text{for } x_{A,0} < x_{B,0}) \tag{4.9}$$

for the ratio of the frequency ν_A emitted by the source and the frequency ν_B measured by the observer Bob. This is the **general formula for the Doppler effect** (in one dimension). One easily recognizes the special cases (4.5) and (4.6).

In the derivation, we have assumed that $x_{A,0} < x_{B,0}$. In the other case, $x_{B,0} < x_{A,0}$, the source has to send the wave in the negative x-direction. We can take this into consideration in formula (4.9) simply by replacing c_S with $-c_S$. The result then is

$$\frac{\nu_B}{\nu_A} = \frac{c_S + v_B}{c_S + v_A} \qquad (\text{for } x_{B,0} < x_{A,0}). \tag{4.10}$$

The whole derivation only works if $|v_B|$ and $|v_A|$ are smaller than c_S. Observer and source must not move faster than the wave in the medium.

There is still an important point to mention concerning formulas (4.9) and (4.10). The frequency shift (or change) depends explicitly on the two velocities v_A of Alice with the source and v_B of the observer Bob (both relative to the medium) and not just on their difference $v_B - v_A$. We will see in Chap. 5 that, for electromagnetic waves (for instance, for light), there is actually no such medium. In this case, the Doppler effect can depend only on the difference $v_B - v_A$ of the velocities of source and observer.[14] If both, v_B and v_A are much smaller than c_S, we get, from (4.9), using the approximation $1/(1+\delta) \approx 1 - \delta$, which is valid for $|\delta| \ll 1$,

$$\frac{\nu_B}{\nu_A} \approx 1 - \frac{v_B - v_A}{c_S} \qquad (\text{for } v_B, v_A \ll c_S \text{ and } x_{A,0} < x_{B,0}).$$

Therefore, even in the presence of a medium, the frequency shift in the case of small velocities depends only on the relative velocity (difference) of observer and source.

Exercise 13: Explain why the derivation above no longer works for $|v_B| > c_S$. What happens if $|v_A|$ is larger than c_S? How is this related to the sonic boom of an airplane?

[14] One should be more precise: the Doppler effect can depend only on the difference $v_B - v_A$ of the velocity v_A of the source when the wave was emitted and the velocity v_B of the observer when the wave was observed.

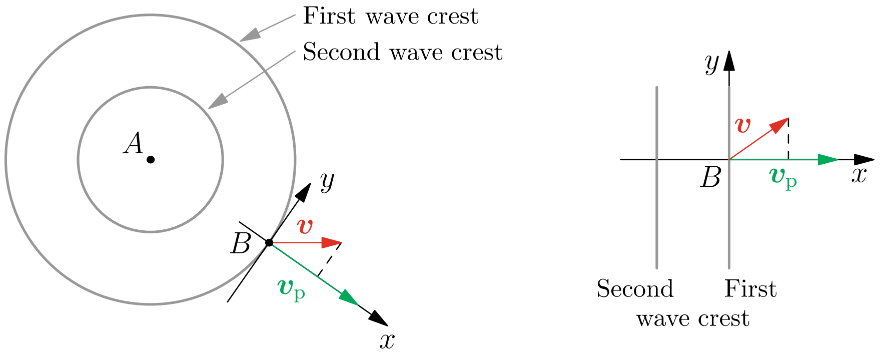


Fig. 4.14 Alice (A) produces spherical waves with wave velocity v_p. Bob (B) moves with velocity $\boldsymbol{v}$ and determines the wave velocity $\boldsymbol{v}'_\mathrm{p}$

Exercise 14: Construct the following $2 \cdot 2 \cdot 2 = 8$ cases:

- Source at rest or observer at rest (relative to medium);
- Observer to the right of the source or to its left;
- Velocity of the source positive or negative.

An example is the case when (a) the observer is at rest (relative to the medium), (b) is located to the left of the source, and (c) the source's velocity is positive. For all these cases, check that the formulas (4.9) and (4.10) correctly predict an increase or a decrease of the frequency ν_B (relative to ν_A).

4.2.5 Aberration

The behavior of the wave velocity (a. k. a. phase velocity). Now, we consider the propagation velocity $\boldsymbol{v}_\mathrm{p}$ of a wave for a moving observer (remember again that, for a moving source, the source velocity does not depend on the velocity of the source).

To measure the wave's velocity, the observer does two things. First, she measures the wavelength λ', i.e., the distance from one wave crest to the next in the direction perpendicular to the wavefront. She gets λ, the wavelength of the wave in the reference frame of the medium: the lengths that she measures do not depend on her velocity relative to the medium. Then, she measures the time T' that elapses from one wave crest passing by her until the next wave crest. This is the wave's period in her reference frame. Then, she calculates the magnitude of the wave velocity $v_\mathrm{p} = \lambda'/\nu'$.

To *determine the behavior of the wave velocity*, she changes the reference frame; we consider the situation in Fig. 4.14, left side. Alice, who is at rest relative to the medium, creates a spherical wave (with wavefronts that are concentric circles). Bob, the observer, at $t = 0$ is located at B and moves with uniform velocity $\boldsymbol{v}$ relative to the medium. We introduce an x-y-coordinate system with the origin in B and with the x-axis pointing in the direction of $\boldsymbol{v}$. The situation is shown again in Fig. 4.14 on the right, where we assume that Bob is far from Alice, and therefore the wavefronts next to him are almost straight lines.

At time $t = 0$, Bob is at the origin of the coordinate system and the first wavefront (with a phase difference of 2π relative to the first one) arrives at his location. When does the second wavefront arrive?

The observer moves with velocity $\boldsymbol{v} = (v_x, 0)$ and the x-coordinate of his position is $x = v_x t$. The second wavefront moves according to $x = v_\mathrm{p} t - \lambda$. The wave's period T', as measured by Bob, is then given by $v_x T' = v_\mathrm{p} T' - \lambda$ or $T' = \lambda/(v_\mathrm{p} - v_x)$. For the wave's velocity, this yields

$$v'_\mathrm{p} = \lambda/T' = v_\mathrm{p} - v_x.$$

We can write this as a vector equation, which eases the comparison to the Galilean addition of velocities (3.5). To do so, we note that the unit vector $\boldsymbol{e}_k$, which is perpendicular to the wavefront, is equal to the unit vector $\boldsymbol{e}_x$ in the x-direction. Therefore, we have $v_x = \boldsymbol{v}\boldsymbol{e}_k$. Thus, we know that the phase velocity is perpendicular to the wavefront, and therefore $\boldsymbol{v}'_\mathrm{p} = v'_\mathrm{p}\boldsymbol{e}_k = (v_\mathrm{p} - v_x)\boldsymbol{e}_k$ and, eventually,

$$\boldsymbol{v}'_\mathrm{p} = \boldsymbol{v}_\mathrm{p} - (\boldsymbol{v}\boldsymbol{e}_k)\boldsymbol{e}_k.$$

The second term $(\boldsymbol{v}\boldsymbol{e}_k)\boldsymbol{e}_k$ is the projection of the velocity $\boldsymbol{v}$ onto $\boldsymbol{e}_k$, the direction perpendicular to the wavefront.

The special cases discussed in Sect. 4.2.2 can easily be seen. When Bob walks perpendicular to the wave's velocity, we have $\boldsymbol{v} \perp \boldsymbol{e}_k$, and therefore $\boldsymbol{v}'_\mathrm{p} = \boldsymbol{v}_\mathrm{p} - (\boldsymbol{v}_p\boldsymbol{e}_k)\boldsymbol{e}_k = \boldsymbol{v}_\mathrm{p}$. If he walks parallel to the wave's velocity, we have $\boldsymbol{v} \parallel \boldsymbol{e}_k$, and therefore $\boldsymbol{v}'_\mathrm{p} = \boldsymbol{v}_\mathrm{p} - (\boldsymbol{v}_p\boldsymbol{e}_k)\boldsymbol{e}_k = \boldsymbol{v}_\mathrm{p} - \boldsymbol{v}$. We confirm that the velocity does not behave (transform) as a velocity. If it did, it would transform according to $\boldsymbol{v}'_\mathrm{p} = \boldsymbol{v}_\mathrm{p} - \boldsymbol{v}$ (wrong!), but this is not the case.

Following all of our discussions on the phase and the group velocity, from here on, we will use $\boldsymbol{c}_\mathrm{W}$ for the group velocity. This means, in particular, that we can use the Galilean addition of velocities for the transformation of $\boldsymbol{c}_\mathrm{W}$.

Direction change with aberration. Let us focus on the aberration angle now, i.e., the difference of the angles under which the different observers see a wave source.

Imagine a wave that, in the rest frame S of its medium, propagates with the velocity $\boldsymbol{c}_\mathrm{W} = c_\mathrm{W}\boldsymbol{e}$, where $\boldsymbol{e}$ is a unit vector that represents the propagation direction. Using (4.4), we calculate the velocity $\boldsymbol{c}_\mathrm{W}{}'$ of the wave for Bob, who moves relative to the medium with the velocity $-\boldsymbol{v}$ (for Bob, the medium moves with the velocity $\boldsymbol{v}$ relative to him). To do so, we introduce the unit vector $\boldsymbol{e}'$, which points in the direction of the velocity $\boldsymbol{c}_\mathrm{W}{}'$, and the direction-dependent magnitude $c_\mathrm{W}{}'(\boldsymbol{e}')$, and write $\boldsymbol{c}_\mathrm{W}{}'(\boldsymbol{e}') = c_\mathrm{W}{}'(\boldsymbol{e}')\boldsymbol{e}'$.

Next, we choose the coordinate systems in the reference frames of the medium and of Bob such that like coordinate axes are parallel and $\boldsymbol{v}$ points in the direction of the x- and x'-axes. By φ, we denote the angle between the x-axis and $\boldsymbol{e}$, and by φ', that between the x'-axis and $\boldsymbol{e}'$ (see Fig. 4.11 on the right side). Then, we get, from (4.4),

$$c_{\mathrm{W}}{}' \begin{pmatrix} \cos\varphi' \\ \sin\varphi' \end{pmatrix} = \begin{pmatrix} v \\ 0 \end{pmatrix} + c_{\mathrm{W}} \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix} = \begin{pmatrix} c_{\mathrm{W}}\cos\varphi + v \\ c_{\mathrm{W}}\sin\varphi \end{pmatrix}. \tag{4.11}$$

To determine $c_{\mathrm{W}}{}'$, we eliminate φ' by adding the squares of the components of the vector equation. It follows that[15]

$$\begin{aligned} c_{\mathrm{W}}{}'^2 &= (c_{\mathrm{W}}\cos\varphi + v)^2 + c_{\mathrm{W}}{}^2\sin^2\varphi \\ &= c_{\mathrm{W}}{}^2 + 2vc_{\mathrm{W}}\cos\varphi + v^2 \\ &= c_{\mathrm{W}}{}^2\left(1 + 2\frac{v}{c_{\mathrm{W}}}\cos\varphi + \frac{v^2}{c_{\mathrm{W}}{}^2}\right) \\ &= c_{\mathrm{W}}{}^2\left(1 + 2\beta_{\mathrm{W}}\cos\varphi + \beta_{\mathrm{W}}^2\right) \end{aligned}$$

with the abbreviation $\beta_{\mathrm{W}} = v/c_{\mathrm{W}}$, and therefore

$$c_{\mathrm{W}}{}'(\varphi) = c_{\mathrm{W}}\sqrt{1 + 2\beta_{\mathrm{W}}\cos\varphi + \beta_{\mathrm{W}}^2}. \tag{4.12}$$

If $\boldsymbol{c}_{\mathrm{W}}$ and $\boldsymbol{v}$ point in the same direction, we have $\cos\varphi = 1$ and, consequently, $c_{\mathrm{W}}{}' = c_{\mathrm{W}} + v$. If the vectors point in opposite directions, we have $c_{\mathrm{W}}{}' = c_{\mathrm{W}} - v$. For all other directions, $c_{\mathrm{W}}{}'$ lies between $c_{\mathrm{W}} - v$ and $c_{\mathrm{W}} + v$. This can also be inferred easily from Fig. 4.11 (right side).

To get an equation for the angle φ' as a function of φ, one can divide the y-component of (4.11) by the x-component. This yields

$$\tan\varphi' = \frac{c_{\mathrm{W}}\sin\varphi}{c_{\mathrm{W}}\cos\varphi + v} = \frac{\sin\varphi}{\cos\varphi + \beta_{\mathrm{W}}}, \tag{4.13}$$

as shown in Fig. 4.15.

For small angles φ, in (4.13), we can use the approximations $\sin\varphi \approx \varphi$ and $\cos\varphi \approx 1$ and get $\tan\varphi' \approx \varphi/(1 + \beta_{\mathrm{W}})$. If φ is small, so is $\varphi/(1 + \beta_{\mathrm{W}})$ and $\tan\varphi'$. Thus, $\tan\varphi' \approx \varphi'$ holds, and we have $\varphi' \approx \varphi/(1 + \beta_{\mathrm{W}})$. From Fig. 4.15, one sees that this is a good approximation that holds for angles smaller than roughly $3\pi/4$.

In the extreme case of $\varphi = \pi$, when $\boldsymbol{e}$ and $\boldsymbol{v}$ point in opposite directions, we get $\tan\varphi' = 0$ or $\varphi' = \pi$ (if $\beta_{\mathrm{W}} \neq 1$).

Another special case is given for $\beta_{\mathrm{W}} = 1$; in that case, with the help of the relation $\sin\varphi/(\cos\varphi + 1) = \tan(\varphi/2)$, we get $\varphi' = \varphi/2$.

Fig. 4.15 shows that the most noticeable aberration happens in the region between $\varphi \approx \pi/2$ and $\varphi \approx 3\pi/4$.

[15] The same can be derived more directly by taking the square of (4.4). We choose the longer way because we will also need the equation for the y-component.

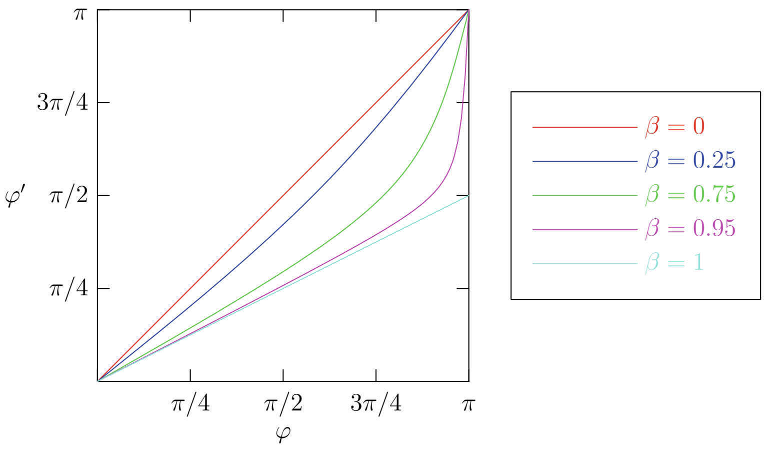


Fig. 4.15 Aberration of the propagation direction of a wave. The angles φ and φ' are defined in Fig. 4.11 on the right side

The boat crossing the river. Everybody has a natural feeling for the change of speed and direction that occurs when the (inertial) reference frame is changed for another one. We experience this, for instance, when we row a boat across a flowing river. Before entering the river, we look at the situation standing on the river bank. The goal is clear: we want to cross the river by the shortest path, which is that perpendicular to the river. Once we are in the river, at rest relative to the river, the situation is different. If we row perpendicular to the river, the flowing water drives us away and we will not cross it by the shortest path.

Let the water flow down the river with velocity $\boldsymbol{v}$. Furthermore, S is the rest frame of the water and S' the rest frame of the river bank. Let the boat move with velocity $\boldsymbol{v}_{\text{Boat}}$ in the rest frame S of the water (see Fig. 4.16). Then, in the rest frame S' of the river bank, the boat moves with velocity $\boldsymbol{v}'_{\text{Boat}} = \boldsymbol{v}_{\text{Boat}} + \boldsymbol{v}$.

We discuss two cases. In the first case, as shown in Fig. 4.16a, the boat moves in S perpendicular to the river bank, $\boldsymbol{v}_{\text{Boat}} \perp \boldsymbol{v}$. Thus, $\varphi = \pi/2$, and therefore $\tan\varphi' = v_{\text{Boat}}/v$. Note that this is consistent with (4.13). For $v_{\text{Boat}} \gg v$, hence, we have $\varphi' \approx \pi/2$, and for $v_{\text{Boat}} = v$, we get $\varphi' = \pi/4$, which is expected.

In the second case, illustrated in Fig. 4.16b, the boat moves such that it reaches the opposite side of the river bank. Then, $\varphi' = \pi/2$, and from the figure, we get $\sin(\varphi - \pi/2) = v/v_{\text{Boat}}$ or $\cos\varphi = -v/v_{\text{Boat}}$. To see that this is consistent with (4.13), we write this formula in the inverted form: $\cot\varphi' = (\cos\varphi + \beta_{\text{W}})/\sin\varphi$. In our case, $\cot\varphi' = 0$. Provided that $\sin\varphi \neq 0$, this requires that $\cos\varphi' = -\beta_{\text{W}}$, which is consistent with our finding from the figure. The equation $\cos\varphi = -v/v_{\text{Boat}}$ only has a solution if $v \leq v_{\text{Boat}}$. If the boat's velocity relative to the water is smaller than the velocity of the river relative to the river bank, it's impossible for the boat to cross the river by the shortest path.

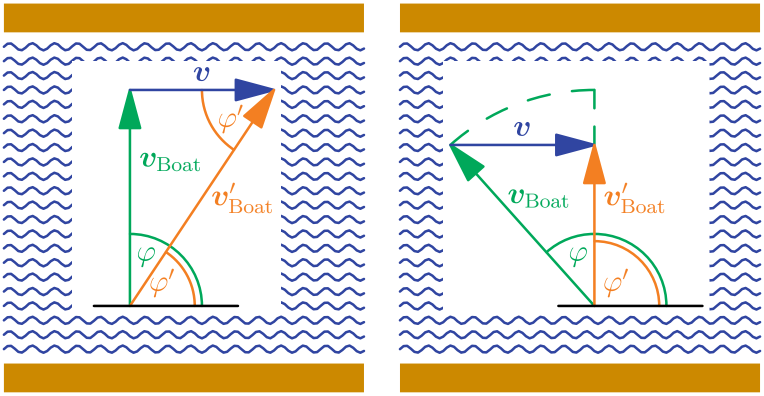

$\boldsymbol{v}$
$\boldsymbol{v}_{\text{Boat}}$
$\boldsymbol{v}'_{\text{Boat}}$
φ
φ'
φ'
$\boldsymbol{v}$
$\boldsymbol{v}_{\text{Boat}}$
$\boldsymbol{v}'_{\text{Boat}}$
φ
φ'

Fig. 4.16 **a** Left: The boat moves in the rest frame of the water, perpendicular to the river's direction. **b** Right: The boat moves such that it reaches the opposite side of the river bank

Note the very loose connection to our discussion on the aberration of waves in media. We assumed (see the discussion in Sect. 4.2.3) that the wavefronts transform as velocity vectors, and therefore reduce our investigation of the effect of aberration to scrutinizing the vector addition formula. In the situation with the moving boat in the river, the description is also based on the velocity vectors addition formula. The fact that there are possibly water waves in the river is not related at all to our investigation of waves.

4.2.6 Waves in Media and the Relativity Principle

An observer who is at rest relative to a homogeneous and isotropic medium uses different laws of Nature than a moving observer. For the former, the propagation velocity of waves in media is independent of the direction, but for the latter, it is not. But this does not imply that the Galilean principle of relativity would not hold for waves in media. The important point here is that the observer *and* the medium have to be transformed to the other inertial frame. This requirement, however, is not a new one. If I move my equipment that I use to make free-fall experiments, to the Moon, it will also yield different results. The Galilean principle of relativity still holds because, if I take the equipment to a different location in space, I also have to take the Earth with me. The Earth is an essential part of my experiment.

4.3 Light as a Wave and the Supposed Luminiferous Aether

4.3.1 *Light is a Wave*

Light is an electromagnetic wave and is described by **Maxwell's equations of electrodynamics**. From these, the **wave equation**[16]

$$\frac{\partial^2}{\partial \boldsymbol{r}^2}\phi(\boldsymbol{r},t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\phi(\boldsymbol{r},t) = 0 \tag{4.14}$$

for light in vacuum follows. In an electromagnetic wave, the electric and magnetic fields oscillate. The quantity ϕ represents one of the components of these fields. The equation above expresses that the propagation velocity of electromagnetic waves is the same at each location and in all directions. The "medium", therefore, is homogeneous and isotropic. The velocity c is the speed of light, which has the value $c = 299{,}792{,}458\,\text{m/s}$.

You are not expected to understand the wave equation. Just remember that it describes the behavior of the electric and magnetic fields in electromagnetic waves, and that light is an electromagnetic wave. The differential quotients $\partial^2/\partial t^2$ and $\partial^2/\partial \boldsymbol{r}^2$ simply denote the second derivatives of the **wave function** $\phi(\boldsymbol{r},t)$ with respect to the time and space coordinates, respectively. If you like, plug the wave $\phi(x,t) = \phi_0 \sin(kx - \omega t)$ into the one-dimensional version of the wave equation $\phi''(x,t) - \ddot{\phi}(x,t)/c^2 = 0$ (the two primes indicate the second derivative with respect to the x-coordinate and the two dots with respect to time) and check what results.[17]

Between classical mechanics and electrodynamics, there is an important difference that is essential for the special theory of relativity. Directly in the fundamental equations of electrodynamics, a (fixed) velocity appears, the speed of light c. There is no such thing in Newton's laws.

4.3.2 *The Medium of the Light Wave?*

We have seen that the equations of classical mechanics (Newton's laws) are valid in all inertial frames and, in particular, that they have the same form in all these reference frames. This is what the Galilean principle of relativity says. The relation

[16] Actually, one should say "the wave equation of electrodynamics" of "the wave equation of light", because there are many other wave equations, such as, for instance, that for sound waves. This chapter, however, is about light, so we can simply keep saying "wave equation" without creating any confusion.

[17] If you made no mistake in the calculation, you will have learned that the wave $\phi(x,t) = \phi_0 \sin(kx - \omega t)$ is a solution to the wave equation *provided that* $\omega/k = c$, which is the dispersion relation for light in vacuum.

between the values of a physical quantity in different inertial frames then is given by the Galilei transformation.

What about the wave equation (4.14) (or Maxwell's equations of electrodynamics)? In which reference frame does it hold? In just one special reference frame or in all inertial frames? This was one of the central questions of physics that ultimately led to the special theory of relativity.

Before Einstein, physicists thought that, in analogy to waves in media, there would also be a medium in which light waves propagate. Exactly as we discussed in Sect. 4.2. And they called this hypothetical medium for electromagnetic waves (or light) the **luminiferous**[18] **aether**. What water is for water waves and the atmosphere for sound waves in the atmosphere, the luminiferous aether would be for light waves. Often, the luminiferous aether was called just the aether.[19]

We will not use the concept of the luminiferous aether here. It is no different from saying that the wave equation is only valid in a special inertial frame and that it has a different form in other reference frames (also in other inertial frames).[20] We call this (hypothetical) special inertial frame *the* **special inertial frame** of electrodynamics.[21] It plays the same role as the luminiferous aether and is the reference frame in which (4.14) holds. But do not get used to this special inertial frame, because, in a few pages, we will see that there is no luminiferous aether (or special inertial frame). Light does not need a medium to exist! But until then, we will keep following the line of thought of classical physics.

We have seen in Sect. 3.4.2 on the Galilei transformation what Bob has to do in order to describe these experiments. He must apply the wave equation in the special inertial frame (which is Alice's reference frame) and then transform the results to his reference frame. Alternatively, he can transform the wave equation to his reference frame and then solve it. He would get a wave equation, where the propagation velocity depends on the direction.

That electrodynamics should be valid in the same form in all inertial frames was an absurd idea to physicists before Einstein. If electrodynamics (and, in consequence, the wave equation) were to hold in all inertial frames in the same way that classical

[18] *luminiferous*: This word comes from the Latin. It is composed of *lumen*, which means "light" and *-ferous*, which comes from the Latin *ferre* and means "to bear, carry, support". So, *luminiferous* means "light-carrying".

[19] In chemistry, "ethers" is a widely used name for a class of organic compounds. They have nothing to do with the luminiferous aether. The word comes from the Latin *aeternus*, which means "eternal". *Aeternus*, on the other hand, comes from the Greek $\alpha\iota\theta\acute{\eta}\rho$, which stands for the blue heaven. So, it means something like "everlasting" or "perpetual".

[20] There is one difference between the luminiferous aether and a special inertial frame, and this difference played a historical role. The medium of a wave – here, the luminiferous aether – does not necessarily have to be rigid. It could be non-homogeneous and/or non-isotropic, and these properties also could depend on time. For the atmosphere, this is exactly the case and leads to effects like the afterglow or the fata morgana, as we explained already. An inertial frame per definition is always rigid. If the luminiferous aether, however, were not rigid, the wave equation (4.14) could not be strictly valid.

[21] The interplay of classical mechanics and electrodynamics shows that this special reference frame indeed has to be an inertial frame.

mechanics does, there would be no special inertial frame and Alice *and* Bob would get the same velocity c for the same light wave. Fig. 1.1 makes this statement clear. Alice is in a inertial frame. Bob moves relative to her with the constant velocity $\boldsymbol{v}$. Now, Alice sends a light beam in Bob's direction. She measures the velocity of the light beam and gets the speed of light c. The light beam passes by Bob, who then measures its velocity as well. Were wave equation (4.14) to be valid in the same form for Bob, he would also get the speed of light c. But this is in contradiction to the Galilean addition of velocities (and, consequently, to the Galilei transformation).

According to the Galilean addition of velocities, the speed of light for Bob should depend on the direction (in the situation shown in Fig. 1.1, he should get $c - v$ for the speed of light). For physicists before Einstein, the wave equation could not be valid for all inertial frames. There was a need for a special inertial frame in which the wave equation is valid in the form of (4.14).

If the wave equation were valid in each inertial frame, then one and the same light beam would have the same velocity c in inertial frames moving relative to each other. This contradicts the Galilean addition of velocities.

4.4 Stellar Aberration

4.4.1 Bradley's Discovery

As an interesting practical example for the aberration of light, we discuss **stellar aberration**. The result that we will work out in this section, however, is not yet completely correct. Later, in Sect. 12.5.3, we will see that the special theory of relativity yields important corrections to the result.

Bradley's discovery: stellar aberration. When observing stars, one finds an aberration effect. This phenomenon was discovered in 1725 by James Bradley[22] and is called **stellar aberration**: if you measure the position of a star, you will find that, in the course of a year, it changes and draws a small ellipse in the heavens.

Clearly, this statement is useless without giving the coordinate system in which the measurements are described. The first thing to note is that the position of a star here is described by the direction in which we see it. Its distance is irrelevant. The same situation prevails when specifying a location on the Earth's surface. One also gives just the direction of the location, as seen from the center of the Earth. In practice, one places an orthogonal coordinate system, the *geographic coordinate system*, with its origin at the center of the Earth in such a way that the z-axis corresponds to the rotation axis of the Earth. The orientation of the coordination system is then fixed by

[22] English physicist, 1692–1762.

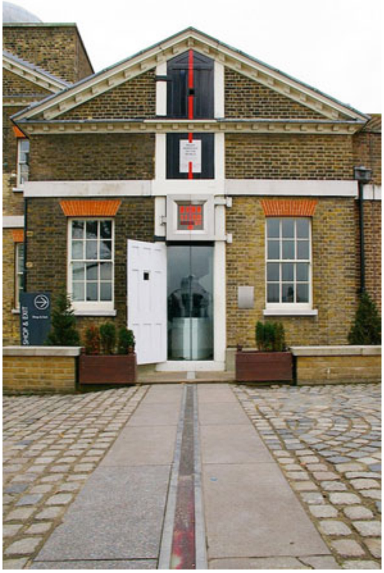


Fig. 4.17 The historic prime meridian, as marked at the Royal Observatory in Greenwich, London

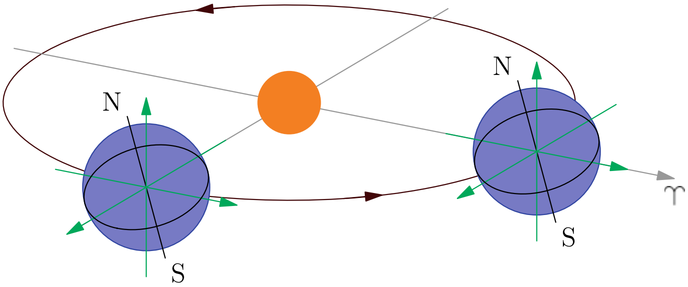


Fig. 4.18 Definition of the ecliptic geocentric coordinate system

defining a point on the Earth that lies on the half plane $y = 0$, $x > 0$. This point is given by the marking of the prime meridian at the Royal Observatory in Greenwich, London (see Fig. 4.17). To specify the location of an arbitrary point P on the Earth's surface, one uses the *polar angle* ϑ, which is the angle between the line segment $\overline{OP}$ and the z-axis and corresponds to the geographic *latitude*, and the *azimuthal angle* φ, which is the angle between the projection of $\overline{OP}$ to the x-y-plane and the x-axis and corresponds to the geographic *longitude*.

To specify the location of a star, one can use the **ecliptic geocentric coordinate system** (egcs). Its origin is at the center of the Earth (that's why it is called "geocentric"), and consequently it moves together with the Earth around the Sun (see Fig. 4.18). The z-axis of the egcs, however, does not coincide with the rotation axis

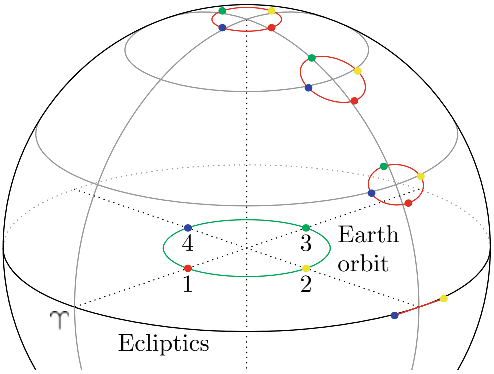


Fig. 4.19 Aberration ellipses

of the Earth, but is perpendicular to the *orbital plane* upon which the center of the Earth moves on its annual journey around the Sun. This orbital plane is called the **ecliptic** (see Fig. 4.19), which explains the other adjective in the name of the egcs. The apparent orbit, which the Sun, as seen from the Earth, described in front of the fixed star sky in the course of a year, thus lies on the x-y-plane of the egcs. The direction of the x-axis is given by the intersection of the equatorial plane of the Earth and the x-y-plane of the egcs and is called the **spring equinox**.[23] The Earth, on March 20, is usually in the position of the spring equinox.[24]

Similar to the geographic coordinate system, the direction of a star is determined by means of two angles. These are the *ecliptic longitude* λ, which corresponds to the geographic longitude, and the *ecliptic latitude* β, which corresponds to the geographic latitude. Due to the way we fixed the x-axis, the coordinate system does not rotate with the Earth around itself. The effect of the Earth's rotation is compensated for in this coordinate system (but not the effect of the orbital motion of the Earth around the Sun, which eventually leads to stellar aberration).[25]

Now, if one measures the direction of stars in the course of a year and plots their coordinates in the egcs, one gets small ellipses, the *aberration ellipses* (see Fig. 4.19). At the poles of the ecliptic, the ellipses become circles, and on the ecliptic, they become line segments. The semi-major axis of the ellipses has the same size of $20.5''$ for all stars ($1''$ is an *arc second* and is equal to $1/3600$th of an angular degree, so $1'' = 0.000278°$).[26] A coin, seen from a distance of 10 km, has a similar

[23] The term "spring equinox" is actually confusing, because it occurs at the beginning of fall for the southern hemisphere. Therefore, it is better to use the term "March equinox".

[24] In some years, it can already be so on March 19 or may not get there until March 21.

[25] The ecliptic is defined by the orbital momentum of the Sun-Earth system and is very stable. The spring equinox is defined by this concept, along with the orbital momentum of the Earth (i.e., the equatorial plane). This is less stable and shows precession and nutation, which is mainly due to the fact that the Earth exchanges orbital momentum with the Sun and the Moon. The influence of these effects on the determination of the stellar aberration, however, is negligible.

[26] There is an effect with similar consequences as stellar aberration, the *stellar parallax*. In that, we also have ellipses, but the size of them depends on the distance of the star from the Earth. The

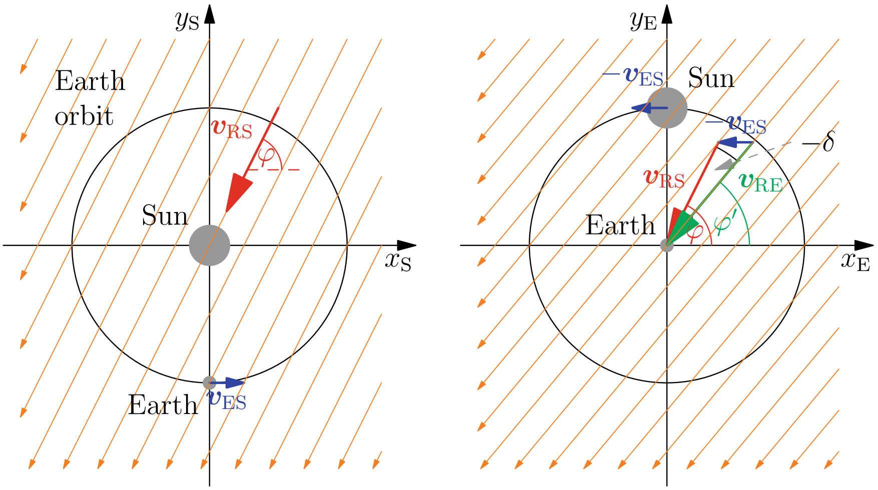


Fig. 4.20 Regarding the discussion of stellar aberration. Left: As seen from the Sun. Right: As seen from the Earth

size. In comparison, the apparent diameter of the Moon in the heavens amounts to about 1860″. Note that stellar aberration can only be observed because it changes in the course of a year! A constant stellar aberration would not be visible.

As indicated in Fig. 4.19 by the digits and the color of the points, the aberration ellipses are not in phase with the position of the Earth in its orbit around the Sun but are rather 90° ahead. This is remarkable, and it shows that the stellar aberration cannot depend on the position of the Earth, but only on its velocity. How can Bradley's observation be explained?

4.4.2 *Bradley's Explanation*

The explanation for the stellar aberration is based on the idea that it is a consequence of the orbital motion of the Earth around the Sun, and therefore, the direction of the light coming from the star, according to (3.5), is different in the respective reference frames of the Earth and the Sun — or the Earth in different positions.[27] The difference in direction, or *aberration angle* δ, only depends on the relative velocity between the observers, i.e., Earth and Sun or the Earth in different positions, and the angle

stellar parallax here is negligible. Even the star closest to the Earth (not considering the Sun), which has the largest stellar parallax, draws an ellipse that is not even 1″ in size.

[27] Note that, while the Doppler effect is an effect between the source and the observer and depends on their relative velocity, stellar aberration is an effect between two observers (the Earth and the Sun or the Earth in different positions) and depends only on the relative velocity of these different observers.

between the direction from which the star light comes[28] and the direction of the relative velocity between the observers.

We will discuss the details now, starting with the perspective of the Sun for which we use the *ecliptic heliocentric coordinate system*(ehcs). Once this is done, we switch to the perspective of the Earth, for which the *ecliptic geocentric coordinate system* (egcs) is useful. The coordinate axes of both coordinate systems are parallel. The only difference between the egcs and the ehcs is that the origin of the former is in the center of the Earth while the origin of the latter is in the center of the Sun.

Perspective of the Sun. Here is the perspective of the Sun. Consider the star R whose light, as seen from the Sun, comes from a direction that lies in the ecliptic, and therefore $\beta = 0$ (its ecliptic latitude vanishes) and has an ecliptic longitude λ_R. The Earth's velocity points towards ecliptic longitude λ_E. The angle φ between the direction of the Earth's velocity and the direction from which the star's light comes and which is the relevant angle to describe stellar aberration is then given by $\varphi = \lambda_R - \lambda_E$.

In Fig. 4.20 on the left side, we show this for the moment when the Earth's velocity points in the x-direction, the direction of the spring equinox ($\lambda = 0$). Then, the angle φ is just the angle between the direction which the Star's light comes from and the x-direction. In this case, the Earth's velocity is given by

$$\boldsymbol{v}_{ES} = v_E \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

where v_E is the magnitude of the orbital velocity of the Earth.[29] The light from the star, relative to the Sun, has the velocity

$$\boldsymbol{v}_{RS} = -c \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix}.$$

We assume that *the Sun is at rest relative to the aether*, so the velocity of light c is the same in all directions.

Perspective of the Earth. Next, we take the perspective of the Earth and transform the velocities to the egcs. To do so, we use the Galilean addition of velocities (3.5), where we identify Alice, Bob, and Claire with the Sun, the Earth, and the star, respectively. Then, $\boldsymbol{v}_{RS} = \boldsymbol{v}_{RE} + \boldsymbol{v}_{ES}$, and therefore

[28] The light coming from the star takes a lot of time to travel to the Sun. Therefore, when the Sun sees the star in a certain direction, the star is possibly already at a position in a completely different direction. The actual direction of the star plays no role.

[29] For simplicity, and because it does not have an effect on stellar aberration, we assume that the Earth's orbit is a circle around the Sun and that the magnitude of the Earth's velocity is constant. Then, the direction of the Earth's position and the Earth's velocity (both relative to the Sun) are perpendicular.

$$\boldsymbol{v}_{RE} = \boldsymbol{v}_{RS} - \boldsymbol{v}_{ES} = \begin{pmatrix} -c\cos\varphi - v_E \\ -c\sin\varphi \end{pmatrix}. \tag{4.15}$$

If we subtract the velocity v_{ES} of the Earth from all velocities in Fig. 4.20 on the left side, we get the situation in Fig. 4.20 on the right side. From the perspective of the Earth, the Sun and the star have an angular distance of $\pi - \varphi'$. The direction, from which the starlight comes, encloses an angle of φ' with the x-axis.[30] From

$$\boldsymbol{v}_{RE} = -c' \begin{pmatrix} \cos\varphi' \\ \sin\varphi' \end{pmatrix},$$

we have

$$\tan\varphi' = \frac{v_{ER,y}}{v_{ER,x}} = \frac{\sin\varphi}{\cos\varphi + \beta_E}, \tag{4.16}$$

where $\beta_E = v_E/c'$ (see also (4.13), where we talked about general waves in media). The orbital velocity of the Earth is about 30 km/s and the speed of light[31] close to 300.000 km/s, and therefore we have $\beta_E = v_E/c' \approx 10^{-4}$. This shows that the effect is small.

If we assume that the Sun is at rest relative to the aether, then the Earth won't be and the light velocity c' will depend on the direction, as shown in Fig. 4.11 on the right side.

The aberration angle for the direction of the star as seen by the Sun and the Earth, respectively, is given by $\delta = \varphi' - \varphi$, and we have

$$\tan(\varphi + \delta) = \frac{\sin\varphi}{\cos\varphi + \beta_E}. \tag{4.17}$$

Using the angle sum formula for the tangent, and after some manipulations, we get a formula for the aberration angle δ:

$$\tan\delta = \frac{-\beta_E \sin\varphi}{1 + \beta_E\cos\varphi}.$$

If, from the perspective of the Sun, the Earth and the star are in the same direction, we have $\varphi = -\pi/2$ and the orbital velocity of the Earth is exactly perpendicular to the direction of the star. This corresponds to the Earth positions 2 or 4 in Fig. 4.19. The stellar aberration then amounts to

[30] In principle, we carried out this calculation already in Sect. 4.2.5. It makes sense, however, to repeat it here. Note that we use a different definition of φ here.

[31] From (4.12), we know that $c' = c'(\varphi) = \sqrt{c^2 + 2v_E c\cos\varphi + v_E^2} = c\sqrt{1 + 2(v_E/c)\cos\varphi + (v_E/c)^2} \approx c\cdot(1 + \beta_E\cos\varphi)$, and therefore, in the expression v_E/c', the difference between c' and c is negligible. This means that we can relax and use v_E/c instead of v_E/c'.

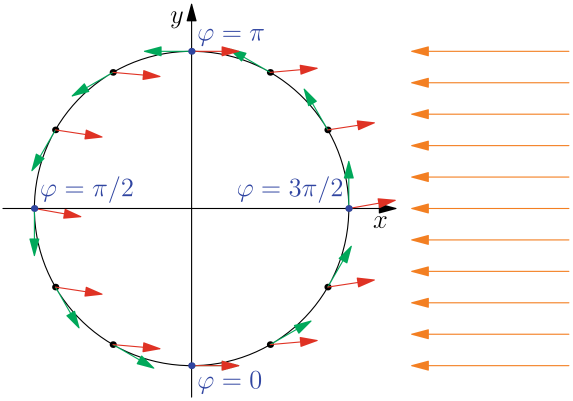


Fig. 4.21 Velocity of the Earth (green vector) and direction (red vectors), in which an observer on Earth sees the light from the star (orange vectors), for different positions of the Earth along its orbit around the Sun

$$\tan\delta=\beta_E \quad \text{or} \quad \delta\approx\beta_E, \tag{4.18}$$

is maximal for all possible angles φ and corresponds to the semi-major axis of the ellipses in Fig. 4.19. With the orbital velocity of the Earth, one gets a value of about $\varphi'=10^{-4}=5.7\times10^{-3°}=20.5''$, which is about one hundredth of the apparent diameter of the Moon – exactly as Bradley measured.

Fig. 4.21 shows how δ changes with φ in the course of a year. If the Earth and the star, as seen from the Sun, have an angular distance of 90°, we have $\varphi=0$ and the orbital velocity of the Earth is parallel to the direction of the star (Earth position 1 or 3 in Fig. 4.19). Then, $\delta=0$. If the Earth, as seen from the Sun, is in the direction of the spring equinox, we have $\varphi=3\pi/2$, the Earth's velocity is perpendicular to the direction from which the Sun's light comes and the aberration angle δ is maximal.

Digression: Sun moves uniformly relative to the aether. In the derivation of (4.18), we assumed that the Sun is at rest relative to the aether. What changes if this is not the case?

Suppose the Sun moves uniformly with β_{SL} relative to the aether (L stands for luminiferous aether) and φ' is the angle under which the star appears for the Sun.

Then, from $\boldsymbol{v}_{RS}=\boldsymbol{v}_{RL}-\boldsymbol{v}_{SL}$, we get

$$\frac{c'}{c}\begin{pmatrix}\cos\varphi'\\ \sin\varphi'\end{pmatrix}=\begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix}+\begin{pmatrix}\beta_{SL}\\ 0\end{pmatrix},$$

where $c'=|\boldsymbol{v}_{RS}|$, or

$$\cot\varphi'=\frac{\cos\varphi+\beta_{SL}}{\sin\varphi}.$$

We consider the case of $\varphi'=\pi/2$, which implies that $\cot\varphi'=0$ or $\cos\varphi=-\beta_{SL}$.

The Earth moves relative to the aether with the velocity β_{EL} and the star appears in the direction φ''.

Then, from $\boldsymbol{v}_{RE} = \boldsymbol{v}_{RL} - \boldsymbol{v}_{EL}$, we get

$$\frac{c''}{c}\begin{pmatrix}\cos\varphi''\\ \sin\varphi''\end{pmatrix} = \begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix} + \begin{pmatrix}\beta_{EL}\\ 0\end{pmatrix},$$

where $c' = |\boldsymbol{v}_{RE}|$, or

$$\cot\varphi'' = \frac{\cos\varphi + \beta_{EL}}{\sin\varphi} = \frac{\beta_{EL} - \beta_{SL}}{\sqrt{1-\cos^2\varphi}} = \frac{\beta_{ES}}{\sqrt{1-\beta_{SL}^2}}.$$

For the aberration angle $\delta := \varphi'' - \varphi'$, we have

$$\cot(\varphi' + \delta) = \tan\delta = \frac{\beta_{ES}}{\sqrt{1-\beta_{SL}^2}}$$

or

$$\delta \approx \beta_{ES}\left(1 + \frac{1}{2}\beta_{SL}^2\right).$$

As long as $\beta_{SL} \ll 1$, this is a very small correction, and in a good approximation, the aberration angle δ resulting from the Earth's motion around the Sun does not depend on the velocity of the Sun relative to the aether.

Digression: Star is outside of the ecliptic. In the explanation of the stellar aberration above, we have assumed that the considered star lies on the ecliptic. We extend our consideration now to the general case. For the calculation, we use the ehcs and rotate the x-y-plane around the z-axis such that the star lies in the x-z-plane. The new coordinates are (x', y', z). Let β_0 be the altitude of the star over the ecliptic. Then, in the new coordinate system,

$$\boldsymbol{v}_{RS} = -c\begin{pmatrix}\cos\beta_0\\ 0\\ \sin\beta_0\end{pmatrix},$$

and the vector of the orbital velocity of the Earth is

$$\boldsymbol{v}_{ES} = v_E\begin{pmatrix}\cos\varphi\\ -\sin\varphi\\ 0\end{pmatrix}.$$

We rotate the coordinate system around the y-axis such that the star becomes located on the ecliptic. This happens by multiplying the vectors with the rotation matrix

$$\hat{R}_y(-\beta_0) = \begin{pmatrix} \cos\beta_0 & 0 & \sin\beta_0 \\ 0 & 1 & 0 \\ -\sin\beta_0 & 0 & \cos\beta_0 \end{pmatrix}$$

and yields

$$\boldsymbol{v}_{RS} = -c \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{v}_{ES} = v_E \begin{pmatrix} -\cos\beta_0 \sin\varphi \\ \cos\varphi \\ \sin\beta_0 \sin\varphi \end{pmatrix}.$$

The projection of $\boldsymbol{v}_{RE} = \boldsymbol{v}_{RS} - \boldsymbol{v}_{ES}$ onto the sphere then gives us

$$\boldsymbol{v}_{RE,\perp} = \beta_E \begin{pmatrix} 0 \\ \cos\varphi \\ \sin\beta_0 \sin\varphi \end{pmatrix}.$$

On the sphere, this corresponds to an ellipse with the semi-major axis β_E in the horizontal direction and the semi-minor axis $\beta_E \sin\beta_0$ in the vertical direction. These are exactly the aberration ellipses shown in Fig. 4.19.

4.4.3 Justification of the Explanation

Bradley's explanation, which well describes the experimental findings, is based on Equation (4.15), the Galilean addition of velocities. He applied it to the velocity of the light ray coming from the star and to the velocity of the Earth relative to the Sun. The challenge now is to justify this explanation.

Particles. Suppose that **the light ray consisted of particles** ("light particles") that behaved according to classical mechanics, like the raindrops in Fig. 3.6. Then, if we were to observe a binary star, the two stars would have different velocities relative to the Earth and, accordingly, the light rays also would have different velocities (we will dive deeper into this in Sect. 5.3.3). This is neither observed nor included in Bradley's explanation. Therefore, an explanation with particles does not work.

Waves in a medium. On the other hand, if **the light ray were a wave in a medium**, the velocities of the light ray coming from the stars would be the same, because the velocity of a wave does not depend on the velocity of the source.

Due to the large distance of the star from the Sun, the wave, however, would be a plane wave, and we had to use the phase velocity (there is no wave packet). The phase velocity, however, does not transform as a velocity, in contradiction to Bradley's explanation. If it were a wave and the group velocity was relevant, Bradley's explanation would be fine. But it is difficult to justify the use of the group velocity. People tried many different explanations but, none really worked.

Another point is the question of the aether. We assumed that the Sun is at rest in the aether. Why should this be the case? If the Sun moves uniformly relative to the aether, we could still explain the found ellipses, but if this velocity were larger, comparable to the speed of light, the predictions would significantly deviate from the observed ellipses.

Enter relativity. These challenges worried physicists for centuries and were eventually resolved by Einstein's relativity. Within the framework of special relativity, both explanations work, that with light particles and that with waves.

The explanation with the light particles works because, in relativity, light particles always travel at the speed of light, which is independent of both the observer and the velocity of the source. This "repairs" the challenge with the binary stars.

And the explanation with the wave works because there will no longer be a difference between the phase and the group velocity. In relativity, the phase velocity of light in vacuum becomes equal to the group velocity and transforms correctly as a velocity. This is due to the fact that simultaneity is no longer absolute in relativity. We will discuss this later.

4.5 Digression: The Transformation of Waves

4.5.1 Transformation of Frequency and Wavevector

What happens with waves when we conduct a Galilei transformation?

Consider the elementary wave (4.1), which, for Alice (and in three dimensions), has the form

$$A(\boldsymbol{r}, t) = A_0 \sin(\boldsymbol{k}\boldsymbol{r} - \omega t). \tag{4.19}$$

For Bob, Alice's wave is still a wave and has the form[32]

$$A(\boldsymbol{r}', t') = A_0 \sin(\boldsymbol{k}'\boldsymbol{r}' - \omega' t'),$$

where $\boldsymbol{r}'$ and t' are given by the Galilei transformation and $\boldsymbol{k}'$ and ω' have to be determined.

We can rephrase our question: what is the transformation that, given $\boldsymbol{k}$ and ω, leads us to $\boldsymbol{k}'$ and ω'?

One thing is clear: at a fixed point P in space and a fixed time t, the wave has a fixed value, and this value is the same for Alice and for Bob. For this reason, we need to have

$$A(\boldsymbol{r}, t) = A(\boldsymbol{r}', t')$$

(where $(\boldsymbol{r}, t)$ are Alice's coordinates of P and $(\boldsymbol{r}', t')$ are the likes for Bob). This is fulfilled if and only if

$$\boldsymbol{k}'\boldsymbol{r}' - \omega' t' = \boldsymbol{k}\boldsymbol{r} - \omega t,$$

i.e., the phase $\varphi = \boldsymbol{k}\boldsymbol{r} - \omega t$ has to be invariant.

[32] We assume that A is a scalar and does not change in the transformation.

The remainder is easy. Using the Galilei transformation, we get

$$\boldsymbol{k}'\boldsymbol{r}' - \omega' t' = \boldsymbol{k}'(\boldsymbol{r} - \boldsymbol{v}t) - \omega' t = \boldsymbol{k}'\boldsymbol{r} - (\omega' + \boldsymbol{k}'\boldsymbol{v})t \overset{!}{=} \boldsymbol{k}\boldsymbol{r} - \omega t.$$

For this to hold, we need

$$\boldsymbol{k}' = \boldsymbol{k}, \tag{4.20}$$

$$\omega' = \omega - \boldsymbol{k}\boldsymbol{v}, \tag{4.21}$$

which is the **Galilei transformation of the frequency and the wavevector** of an elementary wave.

The transformation equation (4.20) for the wave vector is obvious. The wavevector is perpendicular to the wavefronts, and these do not change when changing the inertial frame. We have concluded this already in the discussion of Fig. 4.10. This also means that the wave's wavelength is the same for Alice and Bob.

To understand (4.21), we first restrict ourselves to one dimension.[33] Then, (4.21) becomes $\omega' = \omega - kv$. For waves in which the angular frequency ω and the wavevector k are directly proportional, $\omega = v_\text{p} k$ (see (4.3)). Using this, (4.21) yields $\omega' = \omega - kv = \omega \cdot (1 - v/v_\text{p})$, which is nothing but the Doppler relation (4.9) for the case in which Alice (the source) does not move relative to the medium. In three dimensions, (4.21) is the generalization of the Doppler relation for the case when Alice's velocity relative to the medium and the traveling direction of the wave are not parallel anymore.

4.5.2 *The Dispersion Relation*

When investigating wave phenomena, it is usually easiest to consider waves as linear combinations of elementary waves (4.19).

Such an elementary wave, however, only fulfills the wave equation (i.e., is actually a wave) if the frequency ω and the wavevector $\boldsymbol{k}$ are in a certain relationship to each other. This relationship is the **dispersion relation**

$$\omega = \omega(\boldsymbol{k}).$$

In isotropic media (including the vaccum), the dispersion relation must not depend on the direction. Using $k = |\boldsymbol{k}|$, we have

$$\omega(\boldsymbol{k}) = \omega(k)$$

[33] Note that k can be positive or negative and indicates the traveling direction of the wave, whereas the angular frequency ω is always positive. As a consequence, v_p can also be positive or negative. The speed of light c, however, is positive. Therefore, in one dimension, for a light wave, we have $v_\text{p} = \pm c$.

and can plot the dispersion relation in this case in a two-dimensional ω-k-diagram.

The very special dispersion relation of the form

$$\omega(\boldsymbol{k}) = v_{\mathrm{p}}k$$

(with constant v_{p}) is called **linear**. It is a linear function of the magnitude k of $\boldsymbol{k}$, but not of $\boldsymbol{k}$ itself. An example of a wave with a linear dispersion relation is an electromagnetic wave (including, of course, light) in vacuum. In this case, v_{p} is the speed of light.

Consider an isotropic medium in uniform motion. Its dispersion relation is given by (4.20) and (4.21), and is no longer linear:

$$\omega(\boldsymbol{k}) = \omega_0(k) - \boldsymbol{v}\boldsymbol{k}. \tag{4.22}$$

Here, $\omega_0(k)$ is the medium's dispersion relation in its rest frame. In the one-dimensional case with a linear dispersion relation $\omega_0(k) = v_{\mathrm{p}}|k|$, we have[34]

$$\omega(k) = \begin{cases} -(v_{\mathrm{p}} + v)k & \text{for } k < 0 \text{ (waves traveling to the left)} \\ (v_{\mathrm{p}} - v)k & \text{for } k > 0 \text{ (waves traveling to the right),} \end{cases}$$

which shows that the phase velocity is different in different directions.

The dispersion relation is an important concept and, for instance, allows us to calculate the velocity of the wave.

4.5.3 *The Velocity of a Wave*

What is the velocity of a wave? As we have seen, there are two different definitions. We will not go into details here, but we will explain the consequences.

The phase velocity. We already encountered the *phase velocity in one space dimension* in (4.3). It is basically the velocity with which a wave node moves. Take the wave $A(x, t) = A_0 \sin(kx - \omega t)$, whose phase can be written as $kx - \omega t = k(x - v_{\mathrm{p}}t)$ with $v_{\mathrm{p}} = \omega/k$. Here, $x - v_{\mathrm{p}}t = 0$ or $x = v_{\mathrm{p}}t$ is the equation of motion of a wave node.

The **phase velocity** in three dimensions is defined by[35]

$$\boldsymbol{v}_p := \frac{\omega_{\boldsymbol{k}}}{|\boldsymbol{k}|}\boldsymbol{e}_k \tag{4.23}$$

[34] Note that we have two different definitions of k here. In the two- or three-dimensional case, we use $k := |\boldsymbol{k}|$, while here, in one dimension, we interpret k as the x-component of $\boldsymbol{k}$ – which can be negative.

[35] The two requirements $\boldsymbol{v}_{\mathrm{p}} \parallel \boldsymbol{e}_k$ and $\boldsymbol{k}(\boldsymbol{r} - \boldsymbol{v}_{\mathrm{p}}t) = \boldsymbol{k}\boldsymbol{r} - \omega t$ lead to (4.23).

and always points in the same direction as $\boldsymbol{k}$. In other words, the vector of the phase velocity is always perpendicular to the wavefronts.

Due to the fact that $\boldsymbol{k}' = \boldsymbol{k}$ when we go from Alice's to Bob's inertial frame, the direction of the phase velocity stays the same for a Galilei transformation. The magnitude of the phase velocity, however, changes.

Using (4.20) and (4.21) (in the form $\omega'(\boldsymbol{k}') = \omega(\boldsymbol{k}) - \boldsymbol{v}\boldsymbol{k}$), we can determine the transformation law for the phase velocity:

$$\boldsymbol{v}_\mathrm{p}' \equiv \frac{\omega'(\boldsymbol{k}')}{k'}\boldsymbol{e}_{k'} = \frac{\omega(\boldsymbol{k}) - \boldsymbol{v}\boldsymbol{k}}{k}\boldsymbol{e}_k = v_\mathrm{p}\boldsymbol{e}_k + (\boldsymbol{v}\boldsymbol{e}_k)\boldsymbol{e}_k,$$

which is clearly parallel to $\boldsymbol{k}' = \boldsymbol{k}$. Furthermore, its magnitude is $v_\mathrm{p}' = v_\mathrm{p} + (\boldsymbol{v}\boldsymbol{e}_k)$ and depends on the direction.

Now, $(\boldsymbol{v}\boldsymbol{e}_k)\boldsymbol{e}_k = \boldsymbol{v}_\|$ is the component of $\boldsymbol{v}$ parallel to the wavevector $\boldsymbol{k}$, and we can write this as

$$\boldsymbol{v}_\mathrm{p}' = \boldsymbol{v}_\mathrm{p} - \boldsymbol{v}_\|.$$

Therefore, the phase velocity does not transform according to the Galileian addition of velocities, and thus *the phase velocity is, strictly speaking, not a velocity.*

The group velocity. In Sect. 4.4, when we explained the stellar aberration, we implicitly assumed that the *velocity of a wave* transforms according to the Galileian addition of velocities. Here, we see that, if we take the phase velocity as the velocity of a wave, this is not correct. Fortunately, two circumstances come to our rescue. The first is that there is a further different definition of the *velocity of waves*, the *group velocity*, and this velocity indeed transforms according to the Galileian addition of velocities. And later, in special relativity, the conundrum disappears completely, because the phase velocity of a light wave in special relativity transforms exactly as a velocity (and, again for light, becomes equal to the group velocity).

The **group velocity** is defined by

$$\boldsymbol{v}_\mathrm{g} = \frac{\partial \omega_{\boldsymbol{k}}}{\partial \boldsymbol{k}}. \tag{4.24}$$

In isotropic media, this becomes

$$\boldsymbol{v}_\mathrm{g} = \frac{d\omega(k)}{dk}\boldsymbol{e}_k, \tag{4.25}$$

which is equal to the phase velocity (4.23) for the special case of linear dispersion relations of the form $\omega(\boldsymbol{k}) = v_\mathrm{p}|\boldsymbol{k}|$.

The group velocity arises when one forms a wave packet (see Fig. 4.4(b)) by taking a lot of elementary waves of almost the same frequency. Then, one obtains such a wave packet, and this moves with the group velocity, although its constituting elementary waves move with the phase velocity.

An example in which the phase and group velocities are not equal is given by *deep water waves*. If the amplitude of such waves is much smaller than the wavelength and is in the rest frame of the water, these waves have a non-linear dispersion relation $\omega(k) = \sqrt{gk}$ (g is the standard acceleration due to gravity). We get

$$\begin{aligned} \boldsymbol{v}_\mathrm{g} &= \frac{d\omega}{dk}\boldsymbol{e}_k = \frac{1}{2}\sqrt{\frac{g}{k}}\boldsymbol{e}_k, \\ \boldsymbol{v}_\mathrm{p} &= \frac{\omega}{k}\boldsymbol{e}_k = \sqrt{\frac{g}{k}}\boldsymbol{e}_k, \end{aligned}$$

so $\boldsymbol{v}_\mathrm{g} = \boldsymbol{v}_\mathrm{p}/2$, meaning the group velocity is half as large as the phase velocity.

Again from (4.20) and (4.21) (in the form $\omega'(\boldsymbol{k}') = \omega(\boldsymbol{k}) - \boldsymbol{v}\boldsymbol{k}$), we can determine the transformation law for the group velocity:

$$\boldsymbol{v}'_\mathrm{g} \equiv \frac{\partial\omega'(\boldsymbol{k}')}{\partial\boldsymbol{k}'} = \frac{\partial\omega(\boldsymbol{k})}{\partial\boldsymbol{k}} - \frac{\partial\boldsymbol{v}\boldsymbol{k}}{\partial\boldsymbol{k}} = \boldsymbol{v}_\mathrm{g} - \boldsymbol{v}.$$

This is the transformation law of velocities (the Galileian addition of velocities). The group velocity indeed transforms as a velocity. Therefore, when we interpret the light waves in the explanation of stellar aberration as wave packets, our arguments in Sect. 4.4 hold even in classical physics.

Both velocities, the phase velocity and the group velocity of a wave, can be determined in experiments. One could argue that the different transformation laws for the two velocities gives one the possibility to distinguish different inertial frames. This would violate the Galilean principle of relativity. But this is not the case, because, in classical mechanics, waves always need a medium. And if we go to a different inertial frame without taking the medium with us (as is the case here), we are considering a different physical system. This is analogous to the fact that experiments on the Earth's surface have different results as the same experiments on the Moon's surface. Free-falling objects on the Earth are accelerated six times stronger than on the Moon.

Chapter 5
The Unsuccessful Hunt for the Special Inertial Frame

5.1 First Reflections

According to our argumentation in Sect. 4.3, there should be a special inertial frame in which wave optics is valid and, in particular, in which the light speed in vacuum is the same in all directions. Now, we are interested in the velocity with which we (i.e., the Earth) move relative to this special inertial frame. We will show how this velocity could be measured.

Suppose we can measure this velocity and that the Sun stays at rest regarding the special inertial frame. Then, we could measure our velocity relative to the Sun. We would expect the orbital velocity $v_\mathrm{E} \approx 30\,\mathrm{km/s}$ of the Earth moving around the Sun.

We know already that the speed of light in an inertial frame should depend on both the wave's propagation direction and the velocity $\boldsymbol{v}$ of the inertial frame relative to the special inertial frame. Consequently, to determine our velocity relative to the special inertial frame, we have to measure the speed of light in different directions. We determine the speed of light via the time that a light pulse needs to travel a certain distance. We could do this with a stop watch, but it would be insufficiently sensitive by far. Much better is an interferometric measurement. With an interferometer, one does not measure the length of *one* certain distance (or the time needed for light to travel this distance), but compares the times that light needs to travel two different distances, which typically have almost the same length.

This exact thing has been carried out by Albert A. Michelson, with the subsequent support of Edward W. Morley. The **Michelson-Morley experiment** is one of the most famous experiments in physics. We will discuss it now.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_5

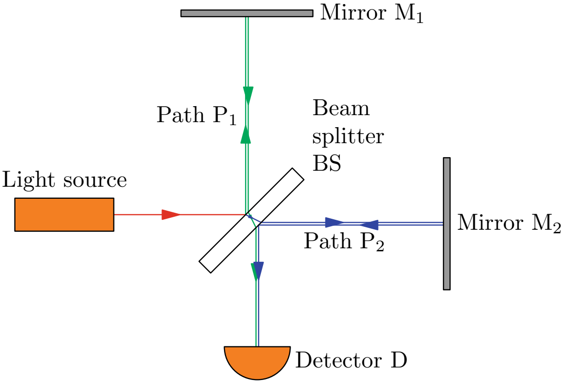


Fig. 5.1 Scheme of the interferometer used by Michelson and Morley (see footnote 1)

5.2 The Experiment by Michelson and Morley and Its Consequences

5.2.1 How It Works

It the preceding chapters, we have learned the tools necessary to put into practice the method discussed in Sect. 5.1 to measure our velocity of us (i.e., that of the Earth) relative to the special inertial frame of wave optics. For this purpose, Michelson and Morley built the device shown in Fig. 5.1, the **Michelson-Morley interferometer**.

The beam path. In this device, a light source produces a monochromatic light beam with the wavelength λ. This light beam impinges on a *beam splitter* BS (e.g., a semitransparent mirror) and becomes split into two partial beams with the same intensity (that's why this particular beam splitter is called a 50–50 beam splitter). One of the partial beams travels along path P_1 to mirror M_1, becomes reflected back to the beam splitter, is transmitted there, and arrives at detector D. The other partial beam travels along path P_2, is reflected at mirror M_2, is further reflected at the beam splitter, and arrives at detector D.[1] The paths P_1 and P_2 are the *interferometer arms*.

Interference. On the path between the beam splitter and the detector, both partial beams interfere. Because both partial beams have the same frequency, wavelength, intensity, and polarization, in front of the detector, we have the situation shown in Fig. 5.2 on the left side.

If one of the partial beams is given by $\phi_1 = a\sin(kx - \omega t)$ and the other by $\phi_2 = a\sin(kx - \omega t + \Delta\varphi)$, where $\Delta\varphi$ is the phase difference, the detector measures the wave

[1] Half of the light beam coming from mirror M_1 is reflected at the beam splitter and travels back to the light source. In the same way, half of the light beam coming from mirror M_2 is transmitted at the beam splitter. This is not relevant to us, as we are interested only in the light arriving at the detector.

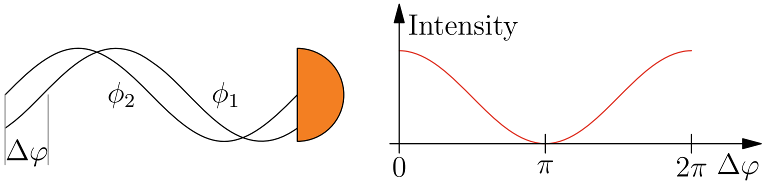
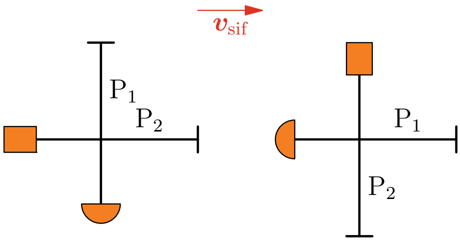


Fig. 5.2 Interference of the partial beams. Left: Situation in front of the detector. Right: Amplitude of the wave in front of the detector in dependence of the phase difference $\Delta\varphi$

Fig. 5.3 Principle of measuring the relative velocity of the Earth relative to the special inertial frame



$$\phi = \phi_1 + \phi_2 = 2a\cos(\Delta\varphi/2)\sin(kx - \omega t + \Delta\varphi/2)$$

with the amplitude $2a\cos(\Delta\varphi/2)$ (see Fig. 5.2, right side) and the intensity proportional to $2a^2\cos^2(\Delta\varphi/2)$. From the intensity measurement, $\Delta\varphi$ can be reconstructed up to a multiple of π. The detector signal is maximal, and one has perfect constructive interference for $\Delta\varphi = 2n\pi$. For $\Delta\varphi = (2n+1)\pi$, however, the detector signal vanishes, and one has perfect destructive interference. In both cases, n is an arbitrary integer number.

Let us assume that both paths have the same length, i.e., the distance from the beam splitter to mirror M_1 is L and the distance from the beam splitter to mirror M_2 is as well. This is the interferometer's *arm length*. A phase difference $\Delta\varphi$ can only then arise from different velocities of the partial beams. Suppose T_1 is the time needed by the light signal to travel along path P_1 (there and back) and T_2 that for path P_2 (also there and back). Then,

$$\Delta\varphi = \omega\Delta T,$$

where $\Delta T = T_1 - T_2$ is the difference in travel time for both partial beams.

Principle of measurement. Using the device of Michelson and Morley, one "measures the interference" in form of the phase difference $\Delta\varphi$ of both partial beams and in this way also the difference ΔT of their traveling times. And how do we determine our velocity relative to the special inertial frame? Suppose, as already mentioned, we knew the velocity $\boldsymbol{v}_{\text{sif}}$ of the special inertial frame relative to us.

Then, we would orient the interferometer such that the path P_2 is parallel to the velocity $\boldsymbol{v}_{\text{sif}}$ (see Fig. 5.3, left side) and the special inertial frame moves away from the light source of the interferometer. What would the device measure?

The speed of light in the rest frame of the interferometer is

$$\boldsymbol{c}' = c\boldsymbol{e} + \boldsymbol{v}_{\text{sif}} \tag{5.1}$$

if $\boldsymbol{v}_{\text{sif}}$ denotes the velocity of the supposed special inertial frame relative to the interferometer [see (4.4)].

First, take **path** P_2, for which $\boldsymbol{c}'$ is parallel to $\boldsymbol{v}_{\text{sif}}$. If the beam travels from the beam splitter BS to the mirror M_1, we have $c'_{\parallel +} = c + v_{\text{sif}}$ (the index $\parallel$ refers to the fact that the path is parallel to the velocity of the supposed special inertial frame and the plus sign indicates the fact that the beam travels outwards). On the return path, we have $c'_{\parallel -} = c - v_{\text{sif}}$. In total, for the traveling time, we get

$$T_\parallel = \frac{L}{c'_{\parallel +}} + \frac{L}{c'_{\parallel -}} = L\left(\frac{1}{c + v_{\text{sif}}} + \frac{1}{c - v_{\text{sif}}}\right) = \frac{2L}{c} \cdot \frac{1}{1 - {\beta_{\text{sif}}}^2}, \quad \beta_{\text{sif}} = v_{\text{sif}}/c. \tag{5.2}$$

If $v_{\text{sif}} = 0$, we get $T_\parallel = 2L/c$, as expected, and for a moving interferometer with $v_{\text{sif}} \neq 0$, $T_\parallel$ is always larger than $2L/c$. Note that we introduced the abbreviation β_{sif} for v_{sif}/c. We will use this abbreviation quite often, as it is very common in relativity. In the theory of relativity, for large velocities, these velocities, in general, are only relevant in proportion to the speed of light. This means that, in the formulas, instead of v_{sif}, v_{sif}/c usually appears.

Now, we look at **path** P_1, for which $\boldsymbol{c}'$ is perpendicular to $\boldsymbol{v}_{\text{sif}}$. These two vectors are the catheti of a right triangle, with $\boldsymbol{c}$ being the hypotenuse. Therefore, the speed of light $c'_\perp$ on path P_1 is $c'^2_\perp = c^2 - v^2_{\text{sif}}$, equal for both the outwards path and the inwards path. In total, for the traveling time, we get

$$T_\perp = \frac{2L}{c'_\perp} = \frac{2L}{c} \cdot \frac{1}{\sqrt{1 - {\beta_{\text{sif}}}^2}}. \tag{5.3}$$

If $v_{\text{sif}} = 0$, one gets $T_\perp = 2L/c$, as expected. And again, for an interferometer with $v_{\text{sif}} \neq 0$, $T_\perp$ is always larger than $2L/c$.

One sees that, for $v_{\text{sif}} \neq 0$, the light for the path parallel to $\boldsymbol{v}_{\text{sif}}$ takes longer than for the path perpendicular to $\boldsymbol{v}_{\text{sif}}$. The travel time difference is

$$\Delta T = T_\parallel - T_\perp = \frac{2L}{c}\left(\frac{1}{1 - {\beta_{\text{sif}}}^2} - \frac{1}{\sqrt{1 - {\beta_{\text{sif}}}^2}}\right). \tag{5.4}$$

Using the approximate formulas $1/(1-x) \approx 1 + x$, and $1/\sqrt{1-x} \approx 1 + x/2$, which are valid for $0 \leq x \ll 1$ one gets the important approximation

$$\Delta T = \frac{2L}{c}\left(\frac{1}{1 - {\beta_{\text{sif}}}^2} - \frac{1}{\sqrt{1 - {\beta_{\text{sif}}}^2}}\right) \approx \frac{L}{c}\left(\frac{v_{\text{sif}}}{c}\right)^2. \tag{5.5}$$

So, the effect is quadratic in v_{sif}/c. In many experiments, the velocities are small. Thus, v_{sif}/c is small and $(v_{\mathrm{sif}}/c)^2$ even smaller, by a lot. In the equation above, one sees that the time difference vanishes for $v_{\mathrm{sif}} = 0$—again, as expected.

After the measurement of ΔT, we rotate the device by 90°. Then, path P_1 is parallel to $\boldsymbol{v}_{\mathrm{sif}}$ and path P_2 perpendicular to it. The quantity ΔT has the same absolute value as in (5.4), but with a negative sign. Therefore, by rotating the device by 90°, the phase difference is doubled. The interferometer, however, has to be very stable mechanically and located in an environment very much free of shocks. Rapid oscillations that change the interferometer's arm length for more than a tenth part of the wavelength (this means less than about 100 nm) already destroy the interference pattern.

A more detailed inspection shows that ΔT in (5.4), for all possible orientations of the interferometer, is that with the largest absolute value if one of the interferometer arms is parallel to $\boldsymbol{v}_{\mathrm{sif}}$. If one measures the traveling time difference ΔT for other angles, one approximately gets a cosine for the dependence of ΔT on the orientation angle.

Now, we calculate $\Delta\varphi$ for the concrete parameter values in the experiment by Michelson and Morley. The arm length was $L = 11$ m, and the gentlemen used the yellow light of a sodium vapor lamp; therefore, $\lambda \approx 590$ nm. Then,

$$\Delta\varphi = \omega\Delta T = \frac{2\pi c}{\lambda}\frac{L}{c}\left(\frac{v}{c}\right)^2 = 2\pi\frac{L}{\lambda}\left(\frac{v}{c}\right)^2 \approx 2\pi \cdot 1.86 \times 10^{-7} \cdot \left(\frac{v}{c}\right)^2 .$$

Suppose now that the Earth travels with its orbital velocity $v_{\mathrm{E}} \approx 30$ km/s relative to the special inertial frame. Then, $v_{\mathrm{E}}/c \approx 10^{-4}$, and therefore $\Delta\varphi \approx 0.75 \cdot \pi$. So, there is an almost complete shift from the constructive to the destructive interference. This would be easily detectable.

A final comment: maybe you noticed that, contrary to what we announced in Sect. 5.1, the Michelson-Morley interferometer does not compare the speed of light in two different mutually orthogonal directions, but rather averages of the "outwards" and the "inwards" speeds of light in these different directions. These averages are called the *two-way* speed of light, in contrast with the *one-way* speed of light.

5.2.2 Result

Michelson carried out the experiment for the first time in 1881 (still without the cooperation of Morley) at the Telegraphenberg in Potsdam, Germany, with the device shown in Fig. 5.4. To his considerable surprise, he was not able to provide evidence for any motion of the Earth relative to the supposed special inertial frame. With an arm length of about 1 m, this first device was about ten times less sensitive than the device that he used later. The expected signal was at the detection limit of the device, and therefore, Michelson was not able to convince the experts.

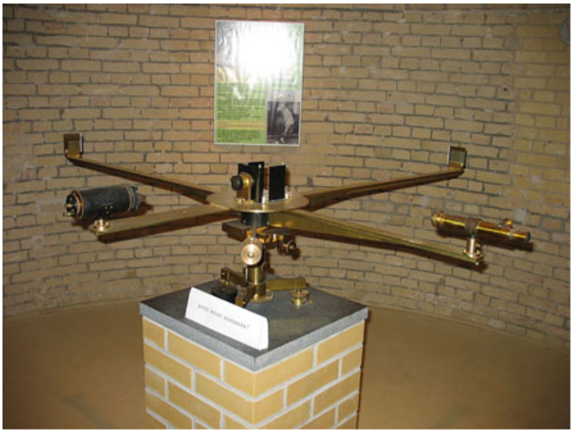


Fig. 5.4 Reconstruction of Michelson's first interferometer, used in Potsdam

Michelson writes [Michelson81]:

> The apparatus […] was placed on a stone pier in the Physical Institute, Berlin. The first observation showed, however, that owing to the extreme sensitiveness of the instrument to vibrations, the work could not be carried on during the day. The experiment was next tried at night. When the mirrors were placed half-way on the arms the fringes were visible, but their position could not be measured till after twelve o'clock, and then only at intervals. When the mirrors were moved out to the ends of the arms, the fringes were only occasionally visible.
>
> It thus appeared that the experiments could not be performed in Berlin, and the apparatus was accordingly removed to the Astrophysicalisches Observatorium in Potsdam. Even here the ordinary stone piers did not suffice, and the apparatus was again transferred, this time to a cellar whose circular walls formed the foundation for the pier of the equatorial.
>
> Here, the fringes under ordinary circumstances were sufficiently quiet to measure, but so extraordinarily sensitive was the instrument that the stamping of the pavement, about 100 meters from the observatory, made the fringes disappear entirely!

When somebody stamps on the floor, even 100 m from the device, the device oscillates, and with it the arm lengths. This also causes the interference pattern to oscillate, which smears it out. A photograph of the interference pattern with an exposure time much smaller than the oscillation's period would still show it. But in that case, one would have a problem with the very low light intensity.

Michelson repeated the experiment in 1889, together with Morley, in Cleveland, Ohio, with an improved interferometer. This device, shown in Fig. 5.5, left side, was much more sensitive than the first one. It had an arm length of 11 m (Michelson achieved the larger arm length by reflecting the light beam several times, see Fig. 5.5, right side). But even the second experiment yielded a null result. The result from the original publication [MM87] is shown in Fig. 5.6. On the x-axis, the orientation of the device is drawn, while on the y-axis, the traveling distance difference for the two interfering paths (in units of the wavelength) is represented. The solid line is the measurement result. For comparison, the dashed line shows the result that was expected for a relative velocity of $v = 30\,\text{km/s}$, but scaled down 8 times! So, the second experiment has confirmed the result of the first one. Michelson and Morley

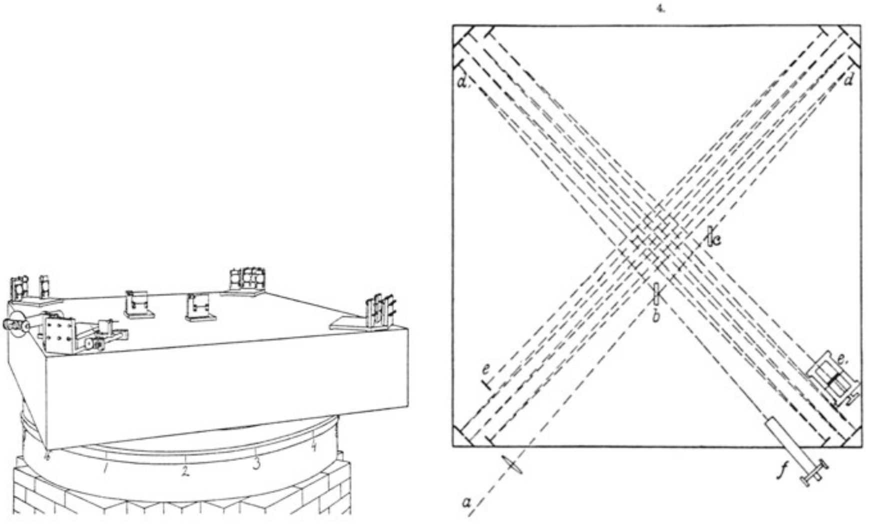


Fig. 5.5 The improved interferometer of Michelson and Morley, as used in Cleveland, Ohio. Left: View from the side. Right: View from above, with the beam path

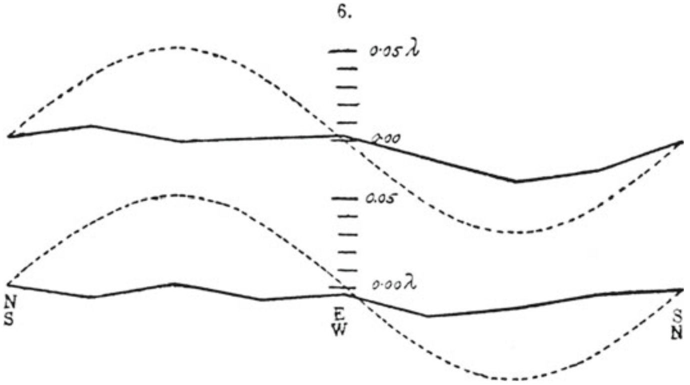


Fig. 5.6 Result of Michelson and Morley's experiment. The path length difference is shown as a function of the orientation. The solid line shows the measurement, while the dashed line shows the expected result *divided by 8*

were not able to measure any velocity of the supposed special inertial frame (or the aether). Within the measurement precision, Michelson and Morley obtained a direction-independent speed of light, exactly as wave optics predicts for the supposed special inertial frame.

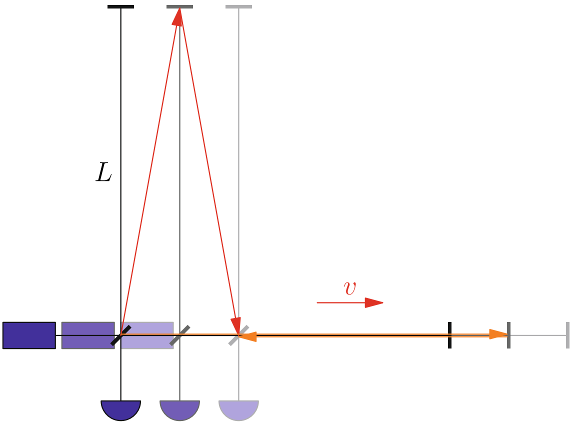


Fig. 5.7 Determination of the light traveling times in the Michelson interferometer, as seen from the special inertial frame (aether reference frame). See Exercise 15 and note that the "red" light pulse and the "orange" light pulse do not arrive at the same time at the respective outer mirrors

The **Michelson-Morley experiment** yields a **null result**. No velocity of the interferometer relative to the supposed special inertial frame can be measured.

A final word: we have already stated that the MM interferometer does not measure the light speed, but rather compares the light speed in the two directions. But even if it were to measure the time that the light pulses need from the beam splitter via the mirrors back to the beam splitter, it would not determine the speed of light, but rather the average of the speed from the beam splitter to the respective mirror and that from the mirror back to the beam splitter. This is called the two-way speed of light. We go deeper into this topic in Sect. 7.7.

Exercise 15: Derive the formula (5.5) for the traveling time difference ΔT in the Michelson interferometer. In contrast to the derivation in the text, do not argue from the point of view of the interferometer's rest frame, but from the special inertial frame. Assume that the interferometer moves with the velocity $\boldsymbol{v}$ such that the direction of the velocity and the direction of the light beam that leaves the light source in the Michelson interferometer coincide. See Fig. 5.7.

5.2.3 *Digression: Arbitrary Orientation of the Interferometer*

In the discussion of the Michelson-Morley experiment, we have assumed that one of the interferometer arms is parallel to the velocity of the supposed special inertial frame (relative to the interferometer's rest frame). This assumption is unrealistic, because we do not know said velocity. So, we have to extend our discussion to the general case of an arbitrary orientation of the interferometer.

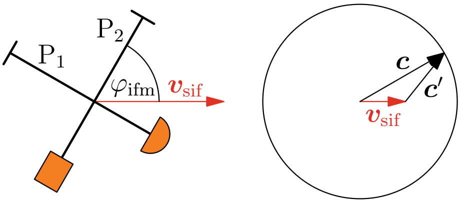


Fig. 5.8 Light velocities in the Michelson-Morley experiment for an arbitrary interferometer orientation

Suppose that the observer (together with the experiment) moves with velocity $\boldsymbol{v}_{\text{sif}}$ relative to the supposed special inertial frame and that the interferometer is oriented such that the direction of path P$_2$ from the beam splitter to mirror M$_2$ of the interferometer and the velocity $\boldsymbol{v}_{\text{sif}}$ makes an angle of φ_{ifm} (see Fig. 5.8, left side).

The velocity c' of the light waves relative to the observer in a direction given by the angle φ' between the given direction and $\boldsymbol{v}_{\text{sif}}$ then follows by squaring $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{v}_{\text{sif}}$ [see (5.1) and the right side of Fig. 5.8] while using $\boldsymbol{c}'\boldsymbol{v}_{\text{sif}} = c'v_{\text{sif}} \cos\varphi'$. This gives us the quadratic equation

$$c'^2 + 2v_{\text{sif}} \cos\varphi' \cdot c' + (v^2 - c^2) = 0$$

with the solution[2]

$$c'(\varphi') = -v_{\text{sif}} \cos\varphi' + \sqrt{c^2 - v_{\text{sif}}^2 \sin^2\varphi'}.$$

The negative sign of the square root is not of interest to us, because it would yield a negative c'. For the angles $\varphi' = 0$ and $\varphi' = \pi$, one gets the minimum and maximum velocities $c - v_{\text{sif}}$ and $c + v_{\text{sif}}$, respectively.

If the interferometer is rotated by the angle φ_{ifm} relative to the direction of the motion of the supposed special inertial frame, one gets

$$\Delta T = T(\varphi_{\text{ifm}}) - T(\varphi_{\text{ifm}} + \pi/2)$$

as a generalization of (5.4).

The time $T(\varphi')$ needed to travel a distance of length L which is oriented in the direction given by φ' relative to $\boldsymbol{v}_{\text{sif}}$, is

[2] Note that we carried out a very similar calculation when we discussed aberration [see (4.11) and the calculation below it].

$$
\begin{aligned}
T(\varphi') &= L\left(\frac{1}{c'(\varphi')} - \frac{1}{c'(\varphi'+\pi)}\right) \\
&= L\left(\frac{1}{-v_{\mathrm{sif}}\cos\varphi' + \sqrt{c^2 - v_{\mathrm{sif}}^2 \sin^2\varphi'}} + \frac{1}{v_{\mathrm{sif}}\cos\varphi' + \sqrt{c^2 - v_{\mathrm{sif}}^2 \sin^2\varphi'}}\right) \\
&= L\frac{2\sqrt{c^2 - v_{\mathrm{sif}}^2 \sin^2\varphi'}}{\sqrt{c^2 - v_{\mathrm{sif}}^2 \sin^2\varphi'}^{\,2} - v_{\mathrm{sif}}^2\cos^2\varphi'} \\
&= \frac{2L}{c^2 - v_{\mathrm{sif}}^2}\sqrt{c^2 - v_{\mathrm{sif}}^2 \sin^2\varphi'}.
\end{aligned}
$$

Eventually, we get

$$
\begin{aligned}
\Delta T &= T(\varphi_{\mathrm{ifm}}) - T(\varphi_{\mathrm{ifm}} + \pi/2) \\
&= \frac{2L}{c^2 - v_{\mathrm{sif}}^2}\left(\sqrt{c^2 - v_{\mathrm{sif}}^2 \sin^2\varphi_{\mathrm{ifm}}} - \sqrt{c^2 - v_{\mathrm{sif}}^2 \cos^2\varphi_{\mathrm{ifm}}}\right).
\end{aligned} \tag{5.6}
$$

The expression in parentheses is $\approx c \cdot (v_{\mathrm{sif}}^2/2c^2)\cos(2\varphi_{\mathrm{ifm}})$. We also use $1/(c^2 - v_{\mathrm{sif}}^2) \approx 1/c^2$. Thus, we get

$$
\Delta T \approx \frac{L}{c}\frac{v_{\mathrm{sif}}^2}{c^2}\cos(2\varphi_{\mathrm{ifm}}).
$$

This is exactly the dashed line in Fig. 5.6. Additionally, for $\varphi_{\mathrm{ifm}} = 0$, one gets the expression (5.5), as expected.

Exercise 16: Show that (5.6) in the special case $\varphi_{\mathrm{ifm}} = 0$ is identical to (5.4).

5.3 Explanation Possibilities

The Michelson-Morley experiment was not able to provide any evidence for a motion of the interferometer relative to the presumed special inertial frame of wave optics—although it was sufficiently sensitive to achieve this easily.

How can we explain this?

We present five possibilities and argue why they would explain the null result. The first three possibilities can be refuted because they cannot explain other important experiments. The fourth can be disregarded for another reason. And the fifth possibility is Einstein's solution to the problem. This solution, detailed in the next chapter, brings us directly to the special theory of relativity.

An important fact to keep in mind: the focus here is on the Michelson-Morley experiment, and so we look for ideas (or theories) that would explain this important experiment. However, it was not only the Michelson-Morley experiment that was unable to bring the aether to light: there were many other experiments, none of which was able to measure the speed of the supposed special inertial frame. So, if we find an idea that can explain the result of the Michelson-Morley experiment, this idea must also explain the results of all these other experiments.

5.3.1 Possibility 1: We Are in the Special Inertial Frame

It could happen that we are in the special inertial frame. This could be the case by chance or the Earth could “drag” the special inertial frame in a certain way. In this case, the speed of light would be the same in all directions at all times and the null result of the Michelson-Morley experiment would be obvious.

Naturally, one has to ask why, in particular, the Earth should rest in the special inertial frame or drag it. Other celestial objects that move with a certain velocity relative to the Earth then should also drag the special inertial frame. And the times when we granted a special role to the Earth are long gone, at least as far back as Copernicus.

In addition to this rational argument, stellar aberration, as discussed in Sect. 4.4, and, in particular, the aberration ellipses in Fig. 4.19 would not be comprehensible if the Earth were at rest relative to the special inertial frame.

Why? The only reason for stellar aberration is the motion of the observer relative to the special inertial frame (or the aether). The apparent direction from which a wave comes from changes only if the observer changes its velocity relative to the special inertial frame. This does not happen if we are in the special inertial frame, so no stellar aberration is expected.[3] In other words: if the Earth were in the special inertial frame, stellar aberration could not be explained. And if the Earth were to move uniformly relative to the special inertial frame, the stellar aberration would not change, and therefore would not be observable.

[3] Actually, in the special inertial frame, the star would move on a circle that compensates the Earth’s motion around the Sun and the observer on Earth would see the *stellar parallax*. This is also perceived as an ellipse, but one much smaller than that for stellar aberration. The stellar parallax has an angle of $\approx D_{\mathrm{ES}}/D_{\mathrm{S}}$ (D_{ES} is the distance from Earth to the Sun and D_{S} that of the star, being largest for the nearest star when it has a semi-major axis of $0.77''$). The stellar aberration, however, is the same for all stars and has a semi-major axis of $20.5''$—much larger. Another difference between the apparent movement of a star due to stellar parallax and the stellar aberration is that both movements are not in phase.

5.3.2 Possibility 2: A Non-homogeneous Luminiferous Aether

In Sect. 4.3.2, we searched for the medium in which light propagates, drew an analogy to the role of water for water waves or that of the atmosphere for sound waves and introduced the luminiferous aether. Then, we concentrated on a rigid luminiferous aether, which is similar to sound waves in a solid, and identified this with a special inertial frame.

We can also think of the luminiferous aether as a liquid or a gas that can be non-homogeneous, non-isotropic, locally moving (flow), or non-stationary (i.e., change its properties with time) or have all these properties at the same time. In this case, even when the medium is locally at rest (no flow), the propagation velocity of light would depend on the direction and the location. The wave would not necessarily propagate on a (straight) line, and among the effects that would then occur is the bending of light rays, as in the fata morgana.

If the luminiferous aether were non-homogeneous and Possibility 2 correct, then, necessarily, wave optics (and therefore Maxwell's electrodynamics) would be wrong, because, according to wave optics, there is a reference frame where light in vacuum propagates on (straight) lines and has the same speed in all directions. If the properties of the luminiferous aether were to be almost homogeneous and isotropic and change only for large distances, this failure of wave optics, however, would possibly be very difficult to detect.

In the 19th century, several aether models were developed. All of them have been ruled out, because they were incapable of explaining the experimental results. Two of these theories are particularly prominent, the aether theory of Augustin Fresnel, developed in 1818, and that of George Stokes, developed in 1844.

Arago's experimental finding and Fresnel's aether drag model. We start with Fresnel's aether model. It assumes that the luminiferous aether is rigid in the vacuum, where this does not matter. Now, the question is: what happens to the luminiferous aether in a dense medium? Would the luminiferous aether flow unperturbed through the medium, as if the medium were not there?

Let us consider a concrete example: a small glass cube (the dense medium) at rest in the luminiferous aether. What would happen to a light ray that impinges perpendicular on the glass cube, which moves relative to the luminiferous aether with velocity v, as shown in Fig. 5.9a? Let us assume that the luminiferous aether moves freely through the dense medium.

If we go to the rest frame of the glass cube (see Fig. 5.9b), we have aberration in the vacuum outside of said cube where the luminiferous aether now moves with velocity $-v$ in the horizontal direction. According to our assumption, the luminiferous aether inside the glass cube also moves with velocity $-v$ in the horizontal direction. Then, due to the fact that the vertical component of the light ray's velocity would be smaller inside the glass cube than outside (the speed of light in vacuum divided by the index of refraction n), we would see a deflection of the light ray *away* from the vertical.

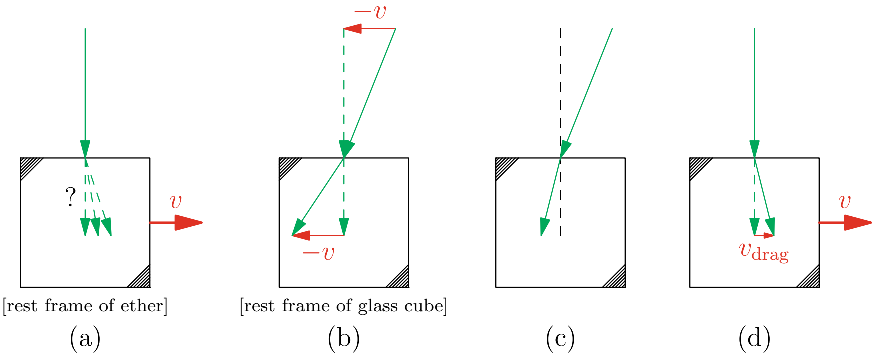


Fig. 5.9 Arago's finding and Fresnel's aether drag coefficient. **a** What happens to a light ray impinging perpendicular to a dense medium that moves with velocity v relative to the luminiferous aether? **b** Assumption: luminiferous aether moves freely through a dense medium. **c** What Arago actually found is consistent with the law of refraction. **d** The luminiferous aether is dragged with velocity v_{drag} by the glass cube that moves with velocity v

François Arago performed an experiment to verify this idea. He took the light of a particular star and carried out the experiment several times a year to make sure that he really captured the situation that the glass cube moves relative to the supposed luminiferous aether. However, he did not observe what was expected when the luminiferous aether moved freely through the glass cube. Instead, he found that, on all occasions, the light ray, upon entering the glass cube, was deflected *toward* the vertical and that it always fulfilled the law of refraction (Snell's law) (see Fig. 5.9c).

How to explain this? If we take the situation with the resting glass cube and the refraction and go back to the system where the luminiferous aether outside the glass cube is at rest and the glass cube moves, we find the situation in Fig. 5.9d. The answer to the question as to what happens to the luminiferous aether inside of the moving glass cube is that the luminiferous aether is *partially dragged* by the glass cube. Instead of the horizontal velocity v, it has the velocity $v_{\text{drag}} = \alpha_n v$ relative to the resting glass cube, where

$$\alpha_n = 1 - \frac{1}{n^2} \tag{5.7}$$

is *Fresnel's aether drag coefficient*. The glass tube, moving with velocity v would drag the luminiferous aether along with itself, with a velocity $\alpha_n v$. If the glass cube were vacuum, we would have $n = 1$ and $\alpha_n = 0$ and there would be no dragging. But the "denser" the optical medium (glass), the larger n and the closer α_n comes to 1. For the limiting case of an infinitely dense medium, we have $\alpha_n = 1$, and the luminiferous aether would be completely dragged and fixed to the dense medium.

Fresnel's formula is fine. In Sect. 10.3, we will calculate the speed of light in moving dense media on the basis of special relativity and show that Fresnel's formula is a good approximation to the relativistically correct formula.

The interpretation with the luminiferous aether, however, is problematic. Due to the fact that the index of refraction n in optically dense media usually depends on the frequency of the light (that's why you see colors in a prism or a rainbow), Fresnel's formula (which is confirmed by Arago's and other experiments) would predict that the velocity with which the glass cube drags the luminiferous aether actually depends on the frequency of the light that impinges on the glass cube. So, there would be a different luminiferous aether for each light frequency and these aethers would be dragged with a different velocity. This was considered highly unlikely and, together with other reasons, *this observation* led to the luminiferous aether being abandoned.

Stokes' aether model. About a quarter of a century after Fresnel, Stokes developed a different aether model. In his model, the Earth moves through the luminiferous aether and completely drags it. What then happens at the Earth's surface is very similar to what you observe on a river bank. The water flows down the river, but very close to the river bank, there is a boundary layer, and the closer the water is to the river bank, the more slowly it flows relative to it. The water clings to the river bank. In the same way, the luminiferous aether, which, far from the Earth's surface, is at rest in the universe (for instance, relative to the Milky Way) clings to the Earth's surface. With this model and the known laws of refraction, Stokes was able to explain stellar aberration. Some years later, however, Lorentz was able to show that Stokes' model contained an untenable assumption. In the end, Stokes' model fell from grace.

5.3.3 *Possibility 3: The Speed of Light is Relative to the Emitter*

If light was not a wave but consisted of particles, the null result would not be particularly surprising (see the example with the firefighters' truck in Sect. 4.2.2). Then, light would not move relative to the special inertial frame with the speed of light, but relative to the emitter. In the Michelson-Morley experiment, the light would have the velocity c relative to the interferometer. Then, it would need the same time for the two interferometer arms (supposing that they have the same length) and the null result would also be obvious. This idea is called the **emitter theory**.

Indeed, the emitter theory is wrong: the speed of light is neither relative to the emitter nor does it depend on the velocity of the emitter. This has been demonstrated with various experiments. Here are two of them.

Alväger experiment. The first generally accepted experiment in this vein was carried out in 1964 by Torsten Alväger and colleagues [Alväger+64] at the European Organization for Nuclear Research (CERN) in Geneva.[4] The physicists shot protons at almost the speed of light through Beryllium nuclei (see Fig. 5.10). In this way, elementary particles called neutral pions (π^0) are produced. These neutral pions

[4] D. Sadeh had already conducted a similar experiment in 1963, but his conclusion was not completely convincing and was not widely accepted by the experts.

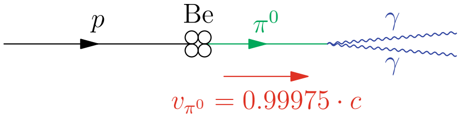


Fig. 5.10 Decaying neutral pions in the Alväger experiment

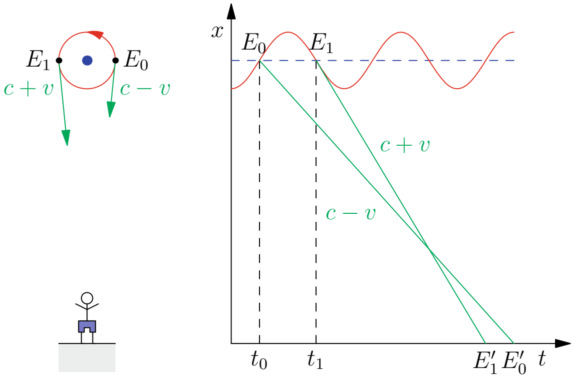


Fig. 5.11 Situation at the binary star. The light ray sent out earlier by the receding star arrives at the observer's location later than the light ray sent out later by the approaching star

have a very large velocity of about $v = 0.99975 \cdot c$ (relative to the laboratory). The important point is that these pions decay into two photons (light particles), which fly approximately in the same direction as the neutral pion. The velocity of the photons, measured in the rest frame of the experiment, within the measurement's precision, was equal to the speed of light c. If the velocity of the photons were to depend on the velocity of the emitter, one would expect them to have almost twice this velocity. Therefore, the Alväger experiment convincingly refuted the idea that the velocity of the light's "corpuscles" depends on the velocity of the emitter.

Brecher's analysis of binary stars. Another argument against the emitter theory is an astronomical one. Imagine a binary star system in which a lighter star orbits a much heavier one (see Fig. 5.11).[5] Let the binary star's orbit be such that the Earth is in its orbital plane. Then, at a certain time t_0, the lighter star will move with a velocity v away from the Earth and, half an orbit later at time t_1, it will move with a velocity v_0 toward the Earth. At time t_0, the light ray that propagates toward the Earth would have the velocity $c - v$, and at time t_1, the velocity $c + v$. The faster light ray would arrive at the Earth earlier than the slower one and the image of the orbit would be distorted or would no longer be recognizable as such. Kenneth Brecher carried out a very detailed analysis of observations of the X-ray binary stars Hercules S-1, Centaurus X-3 and SMC X-1, but noticed no such effect [Brecher77].

[5] To be more precise, both orbit around their common center of mass.

5.3.4 Possibility 4: Lorentz-FitzGerald Contraction and Lorentz's Ether Theory

In the last decade of the 19th century, a not exactly obvious solution to the problem of the null-result of the Michelson-Morley experiment was presented. It was developed in a qualitative way in 1889 by George FitzGerald, and a bit later by Hendrik Antoon Lorentz and Joseph Larmor, quantitatively.

Their hypothesis, called the **Lorentz-FitzGerald contraction**, states that all objects contract in the direction of their motion relative to the special inertial frame by a factor of $\alpha(v) = \sqrt{1 - v^2/c^2}$. For classical mechanics, this would not imply any observable change, because all objects and measurement rods would contract in the same way. Measuring the length of an object with a meter stick, one would get the same result as without the Lorentz-FitzGerald contraction. But light would not be subject to the contraction. For the case of the Michelson-Morley interferometer, instead of (5.2), with $L' = \alpha(v)L = L \cdot \sqrt{1 - \beta^2}$, one would get

$$T_{\parallel} = \frac{2L'}{c}\frac{1}{1-\beta^2} = \frac{2}{c} \cdot (L\sqrt{1-\beta^2})\frac{1}{1-\beta^2} = \frac{2L}{c}\frac{1}{\sqrt{1-\beta^2}} = T_{\perp}$$

and the phase difference $\Delta\varphi$ in the interferometer would be always zero. Based on that, we would expect a null result in the Michelson-Morley experiment. We discuss this in more detail in Sect. 12.6.

So, the result of the Michelson-Morley experiment could be explained with the aether and the hypothesis of FitzGerald and Lorentz. But other effects, like stellar aberration (see Sect. 4.4), the (relativistic) Doppler effect (see Sect. 9.6) or the result of the Fizeau experiment (see Sect. 10.3), cannot be explained on this basis.

Lorentz, however, succeeded in making further modifications, which form **Lorentz's ether theory**. This theory is able to explain all these experiments. One of these modifications is the introduction of an *apparent time* (called *local time* for inertial frames that are different from the aether, the special inertial frame), in addition to the *true time* shown by clocks that rest relative to the aether. An interesting conclusion from Lorentz's ether theory was that, by construction, the aether could never be discovered. So, the theory introduces a physical concept just to prove later that there's no way to discover it!

Einstein's theory is also able to explain all these experiments. But due to the fact that Einstein's theory, from a logical point of view, is much easier than Lorentz's ether theory, the former has come to be preferred. This follows a principle of the philosophy of science called *Occam's razor*. It states that, out of two alternative explanations of a certain phenomenon, one should prefer that that has fewer hypotheses and is conceptually easier.

Nevertheless, this shows that Einstein, to use Newton's words, stood on the shoulders of giants. Of these giants, Henri Poincaré was the one who, in 1905, came closest to the special theory of relativity (see Sect. 6.1). The relevance of the principle of relativity was clear to him, and he stated that there was no possibility of identifying the aether, the special inertial frame. But even he never abandoned the aether. And he thought that only the clocks that rest relative to the aether show the “true time”. The clocks that move uniformly relative to the aether, however, according to him, show an “apparent time”, which would depend on the applied synchronization method.

Einstein then carried out the revolution and got rid of the aether. He took the principle of relativity quite seriously and showed that, solely from this principle and the principle of the absolute speed of light, the special theory of relativity follows. While Lorentz’s ether theory in its final form also can explain the physics at large velocities, it is conceptually more complicated and, in particular, must live with the claim that the aether exists but that we, *by construction*, never will be able to prove this. This is very unsatisfactory.

5.3.5 Possibility 5: There Is No Special Inertial Frame

This is Einstein’s solution to the puzzle. Einstein made the hypothesis that there is no special inertial frame (and, therefore, no luminiferous aether). All inertial frames are equal and light in each inertial frame travels at the speed of light c. Wave optics (or electrodynamics) is valid in all inertial frames. This hypothesis not only explains the null result in the Michelson-Morley experiment, it also makes things much more simple. The effect is similar to that which occurred when the geocentric model with the Earth at the center of the Universe, with all of its complications, such as the epicycles, which were needed to describe the motion of the planets, was replaced by Copernicus’ heliocentric model.

But more of this will be addressed in the remainder of the book.

5.4 Summary

In Sect. 4.3, we argued that there should be a special inertial frame in which wave optics is valid and in which the light propagates in all directions with the same velocity. It is an obvious goal to identify this special inertial frame by measuring its velocity relative to the observer. This was performed by Michelson and Morley, and ended in a null result—although the interferometer was well designed to find the answer to this quest. We discussed several possibilities to explain this null result. Three have been ruled out by further experiments and a fourth one (Lorentz’s ether theory) has been discarded because of its conceptual complexity. The fifth one, by Einstein, states that there is no special inertial frame (or aether) and leads directly to his special theory of relativity.

Chapter 6
Einstein's Solution: The Special Theory of Relativity (SR)

As an explanation for the unexpected result of a non-identifiable special inertial frame in the Michelson-Morley experiment, we discussed Einstein's solution: *there is no special inertial frame!* (see Sect. 5.3.5). But before going into detail with this idea, we go one step back.

6.1 Einstein's Two Principles and Their Consequences

Einstein originally argued from a different point of view, as did we in Sect. 5.3.5. He started with the observation that classical mechanics *in the same form* is valid in all inertial frames. We discussed this property of classical mechanics in Sect. 3.4 and called it Galilei's principle of relativity. It states that, with experiments that are described by classical mechanics, you cannot distinguish two inertial frames. Einstein noticed that this is a very elegant and useful property of Nature. But he asked himself why this only holds for classical mechanics and not for electrodynamics (which is the mother of wave optics). According to his opinion, the principle of relativity should hold for electrodynamics as well. Two inertial frames should be indistinguishable with mechanical *and electrodynamical* experiments. This extension of Galilei's principle of relativity, postulated by Einstein, is commonly called *Einstein's principle of relativity*[1] and is often referred to simply as the *principle of relativity*.[2]

[1] We use the abbreviation EPR for Einstein's principle of relativity, advising the reader that EPR usually stands for the Einstein-Podolsky-Rosen paradox, which is a completely different thing. It contains, in some sense, the essentials of quantum physics.

[2] Note that this does not entail that classical mechanics was correct. It just says that a correct mechanics must have the property of having the same form in all inertial frames. The same applies to electrodynamics and other physical theories.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_6

Einstein's principle of relativity (EPR): Both mechanical and electrodynamical phenomena[3] appear in the same way in all inertial frames.

What are the consequences of Einstein's principle of relativity? First, there would be no special inertial frame for electrodynamics (or wave optics). In Sect. 4.3, we argued that, in the special inertial frame (and only there), light propagates in the same way in all directions. According to Einstein's principle of relativity, this statement would hold in all inertial frames. In all inertial frames, light would have the same velocity c in all directions. Einstein's principle of relativity explains the zero-result of the Michelson-Morley experiment without any hassle. We formulate the essence of this idea as a principle:

Principle of the absolute speed of light (PASL): Independently of the inertial frame, light propagates (in vacuum) in all directions with the same velocity, the *absolute*[4] speed of light c.

The principle of the absolute speed of light elevates the *speed of light* in vacuum c to a fundamental physical constant.[5,6]

As already mentioned, with the PASL in place, the explanation of the zero-result of the Michelson-Morley experiment comes without any cost. But it also has a side that seems ugly at first sight. In Sect. 4.3, we argued that there should be a special inertial frame for wave optics. The reason was the Galilean addition of velocities.

Let us recall the circumstances of Einstein's principle of relativity (see Fig. 1.1). The inertial observers Alice and Bob move relative to each other with the velocity v. A light pulse passes by both. If wave optics is valid for both, **Alice measures exactly the same velocity of the light pulse as Bob**, the absolute speed of light. It is this observation that is confirmed by the zero-result in the Michelson-Morley experiment.

Again due to the discussion in Sect. 4.3, this can only be correct if the Galilean addition of velocities does not hold anymore (at least for large velocities). There must be a new law for the addition of velocities, which, being inspired by (3.5), should have the form

$$\boldsymbol{v}_{\mathrm{AC}} = \boldsymbol{v}_{\mathrm{AB}} \oplus \boldsymbol{v}_{\mathrm{BC}}.$$

[3] Electrodynamical phenomena include light phenomena.

[4] The adjective *absolute* brings to our attention that the speed of light does not depend on the inertial frame. In other words: it is not *relative* (to the inertial frame).

[5] Actually, Einstein formulated the principle of the absolute speed of light in a different way, namely that (in the observer's inertial frame) the speed of light is independent of the velocity of the light source. Together with the principle of relativity, this is tantamount to our formulation.

[6] The principle of the absolute speed of light depends in a sophisticated way, on the way in which clocks are synchronized (see Sect. 7.7).

We will find out the exact meaning of the binary operation $\oplus$ in Chap. 10 and call it the **Lorentzian addition of velocities**. We already have two hints: due to the fact that the Galilean addition of velocities is correct for small velocities, the operation $\oplus$ should become a simple addition within this limit. Additionally, the discussion above (see Fig. 1.1) has shown that, when adding an arbitrary velocity v and the speed of light c, the speed of light must result: $v \oplus c = c$.

The Galilean addition of velocities becomes wrong for large velocities. According to (3.9), the exact same then holds for the Galilei transformation. The consequence is that, according to (3.4), *classical mechanics is no longer applicable for large velocities*. Exactly as we have seen in the Bertozzi experiment (Chap. 2). To summarize: due to Einstein's two principles, the GAV cannot be valid anymore. Consequently, the GT is wrong as well (at least, for the considered large velocities). And this implies that either the GPR or classical mechanics is wrong. As the GPR is a special case of the EPR, it is classical mechanics that must be wrong.

As a consequence, we have to replace classical mechanics with a *relativistic mechanics*. This was done by Einstein in 1905 [Einstein05a]. Relativistic mechanics is a part of his **special theory of relativity (SR)**. This theory, in particular, must yield the correct law for the addition of velocities, even at large velocities.

Einstein has shown that, to understand the consequences of relativistic physics to space and time (the "kinematics"), nothing more than the two above principles are needed: **Einstein's principle of relativity** and the **principle of the absolute speed of light**. This is not sufficient, however, to construct the relativistic replacement for Newton's laws (the "dynamics"). To be able to achieve this, one additionally has to require that the relativistic laws for small velocities become identical to the laws of classical mechanics.

6.2 The Relevance of the Principle of Relativity

Einstein's principle of relativity looks innocent, but actually, it is quite powerful and has ample consequences for physics. There are **symmetry principles**[7] with similar far-reaching consequences:

- No matter *where* one carries out an experiment, the result is always the same. This is called the **homogeneity of space**.
- No matter *when* one carries out an experiment, the result is always the same. This is called the **homogeneity of time**.
- No matter *in which direction* one carries out an experiment, the result is always the same. This is called the **isotropy of space**.

[7] An object is **symmetric** regarding a set of operations if these operations leave the object unchanged or invariant. A circle is invariant regarding rotations about its center. An equilateral triangle is invariant regarding rotations of a multiple of 120° around its center. The distance $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$ of two points $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$ in Euclidean space is invariant regarding arbitrary rotations and translations.

These principles are important because of their consequences. From the homogeneity of space[8] follows the *conservation of momentum* (i.e., Newton's third law). From the homogeneity of time follows the *conservation of energy*. And from the isotropy of space follows the *conservation of angular momentum*. Symmetry principles and conservation laws are very closely related.

The principle of relativity extends this list:

- No matter *in which inertial frame* one carries out an experiment, the result is always the same. This is **Einstein's principle of relativity**.

From the independence of the particular inertial frame follows the conservation of a quantity related to the center-of-mass motion.

But now, let's roll up our sleeves. We have to work out Einstein's theory, and will do so after a short digression on the measurement of the speed of light.

6.3 Digression: Measuring the Speed of Light

Measuring the speed of light has always been an important task in physics. The first measurements were carried out by Galilei in about 1620. Due to the fact that light's speed is so fast, Galilei could only give a lower boundary to it. During the following centuries, light's speed was measured many times, with different methods, and an ever increasing precision.

In 1905, when Einstein's principle of the absolute speed of light elevated the speed of light to the rank of a fundamental physical constant, the determination of its value became even more important. Nowadays, the value of the speed of light is fixed, because the meter is defined via the second. The precision of the speed-of-light measurement needed to allow for this definition, however, was only achieved in 1973 at the National Bureau of Standards (NBS)[9] (see Sect. 6.3.5). So, the measurement of the speed of light has kept physicists busy for 350 years! We will mark some of the milestones along this path.

The first value that was roughly correct was measured by Ole Rømer in 1676, using observations of the orbital period of Jupiter's moons. To be able to carry out his method, one needs to know the difference between the maximum and the minimum distance between Jupiter and the Earth, in other words, the *diameter of the Earth's orbit around the Sun*. Therefore, we start by discussing how some astronomical quantities of the solar system are determined.

[8] And an additional assumption about the form of physical theories. The relation between this type of symmetry (continuous symmetries) and conservation laws is called *Noether's theorem*.

[9] Since 1988, the NBS has been called the *National Institute for Standards and Technologies* (NIST). It is located in Boulder, Colorado, USA. This institute in the US has similar tasks as the Physikalisch Technische Bundesanstalt (PTB) in Braunschweig, Germany.

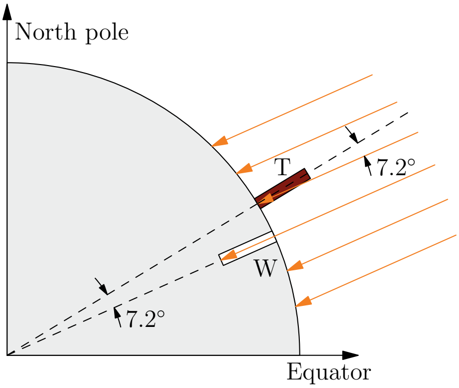


Fig. 6.1 Determination of the Earth's circumference according to Eratosthenes. W indicates the water well in Syene (Assuan) and T the tower in Alexandria

6.3.1 Determining Distances and Sizes in the Solar System

Earth's radius. To determine the diameter of the Earth's orbit (i.e., twice the average distance D_{ES} between the Earth and the Sun), which is needed in Rømer's method to measure the speed of light, we start with the **Earth's radius** R_{E}. This quantity had already been determined by Eratosthenes in about 200 BCE. His method (see Fig. 6.1) is based on the observation that, on a certain day of the year at noon, the Sunlight falls exactly vertically into a water well W in the City of Syene (now Assuan) in Egypt. On the same day in Alexandria, also at noon, the Sun is about 7.2° away from the vertical position, marked by the tower T. This means that the difference in geographic latitude of both cities is exactly this value.[10] Both cities have roughly the same longitude. Therefore, the distance between the cities corresponds to about a fiftieth ($= 7.2°/360°$) of the Earth's circumference. Measuring the distance between the cities, Eratosthenes got 5000 stadia, therefore, the circumference of the Earth would be 250,000 stadia. Unfortunately, the exact length of a stadion is not known, but taking the actual distance between the cities of 835 km, one gets 41,750 km for the Earth's circumference. This is impressively close to the actual value of 40,075 km at the equator. The modern value[11] for the circumference is 40.030 km and the radius is $R_{\mathrm{E}} = 6.371$ km.

The next steps on way to the distance D_{ES} between the Earth and the Sun were carried out in about 250 BCE by Aristarchus of Samos. With his method, one gets D_{ES}, but also several other distances in the solar system.[12] And this is only possible by combining the results of several observations.

[10] Here, one has to assume that the Sun is far away from the Earth. Much farther than the distance between the two cities.

[11] The Earth is not a sphere, therefore, the definition of its radius and circumference is, to some degree, arbitrary.

[12] Aristarchus got all these quantities in relation to the Earth's radius R_{E}. The latter was not yet know at this time, as Eratosthenes' discovery didn't come about until 50 years later.

Apparent size of the Sun and Moon. We start with the first observation, which is relatively simple: the apparent size of the Sun and the Moon in the heavens is almost equal. This can best be seen during a solar eclipse, when the Moon almost completely covers the Sun. Figure 6.2 shows the situation, and one easily sees the relation (R_{E} is less than 2% of D_{EM}, so it usually neglected in $D_{\mathrm{EM}} - R_{\mathrm{E}}$ and in $D_{\mathrm{ES}} - R_{\mathrm{E}}$)

$$\frac{R_{\mathrm{S}}}{D_{\mathrm{ES}}} = \frac{R_{\mathrm{M}}}{D_{\mathrm{EM}}}, \tag{6.1}$$

$$\frac{R_{\mathrm{S}}}{D_{\mathrm{ES}}} = \tan\frac{\alpha}{2}, \tag{6.2}$$

which holds to a very good approximation and where α is the apparent size, which is measured to be about $\alpha = 0.53°$.

Angle between the Sun and Moon at half moon. Aristarchus used an interesting method to get a relation between the distance D_{ES} from the Earth to the Sun and that D_{EM} to the Moon. In Fig. 6.2, right side, the constellation of the Sun, the Earth, and the Moon is shown at new moon, half moon, and full moon. At new moon, the Moon is in between the Sun and the Earth (usually only approximately, otherwise a solar eclipse would result). At full moon, the Moon again lies on the same line as the Earth and the Sun. This time, however, the Earth is sitting in the middle (again approximately, otherwise a lunar eclipse results). At half moon, the Moon, as seen from the Earth, is not exactly 90° away from the new-moon position, but a little bit less: the Earth-Moon-Sun angle must be a right angle.

Measured from new moon, the position of the Moon is therefore smaller than 90°. Let us call the difference β. This angle can be determined by comparing the duration from new moon to half moon to that from half moon to full moon. The difference between these durations is only half an hour, therefore, $\beta = 0.5\,\mathrm{h}/(1\ \mathrm{month}) \approx 1/390$. The angle β is also the angle between the direction of the Earth and the Moon at half moon and as seen from the Sun. Therefore, we have

$$\frac{D_{\mathrm{EM}}}{D_{\mathrm{ES}}} = \sin\beta \approx \frac{1}{390}. \tag{6.3}$$

The Sun is almost 400 times farther from the Earth than the Moon! Because of (6.1), we also have

$$\frac{R_{\mathrm{M}}}{R_{\mathrm{S}}} = \frac{D_{\mathrm{EM}}}{D_{\mathrm{ES}}} \approx \frac{1}{390}. \tag{6.4}$$

The weakness of Aristarchus' method is that the angle β is very difficult to determine. Aristarchus got about 3°, which is far too large. The consequence was a far too low estimate of the size of the Sun: Aristarchus inferred that the Sun was between 18 and 20 times farther away from the Earth than the Moon. The actual factor is 390! Nevertheless, Aristarchus' method is correct.

Now, we have the four unknowns R_{S}, D_{ES}, R_{M}, D_{EM} (R_{E} is known) and three Eqs. (6.1)–(6.3). Too little to determine the unknowns.

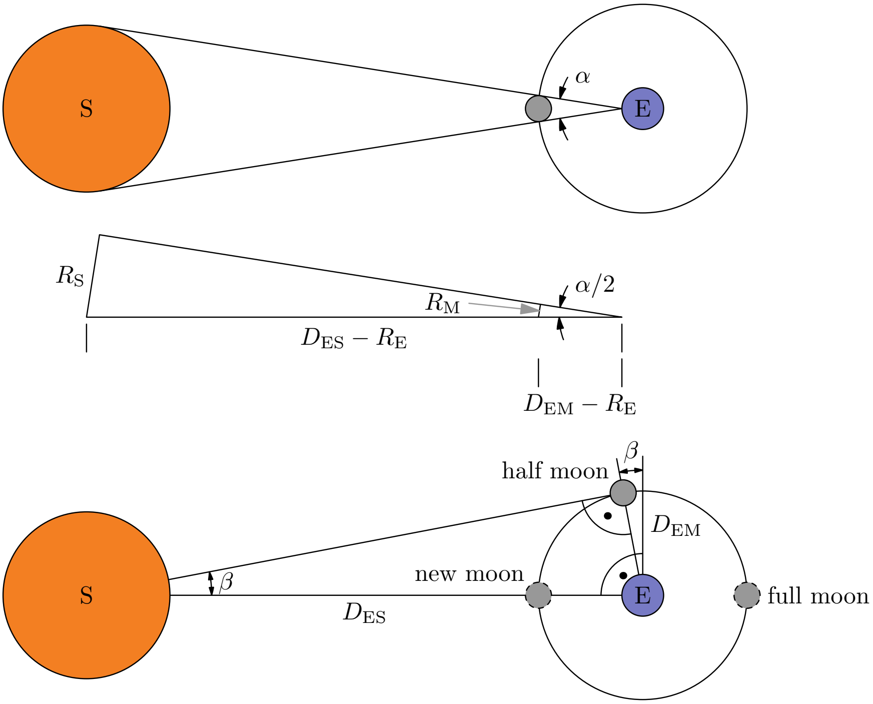


Fig. 6.2 Top: Solar eclipse and apparent sizes of the Sun and Moon. Bottom: Determination of the ratio between the distance D_{ES} from the Earth to the Sun and D_{EM} from the Earth to the Moon via the lunar phases

The lunar eclipse. The last needed piece of information comes from the observation of a lunar eclipse. Figure 6.3, top left, shows the situation. The Earth, in the sunlight, produces a shadow. First, there is a total shadow, the *umbra*. Looking from any point in the umbra, the Sun is completely covered by the Earth. Then, there is a partial shadow, the *penumbra*, in which the Sun is partially visible. If you move within the shadow of the Earth from A to B (see Fig. 6.3, top right), it is completely dark. At B, the Sun starts to appear, and it becomes lighter. On the way from B to C, it becomes gradually lighter until, at C, you see the full Sun.

The idea now is that the ratio between the Moon's radius R_M and the radius R_U of the umbra can be determined by observing the Moon moving through the umbra. To do so, one can measure the time T_{DE} that the Moon takes from point D (top right of Fig. 6.3) to point E, where it covers a distance of $2R_M$, to the time T_{DF} from D to F, where it covers $2R_U$, and obtain $R_U/R_M = T_{AC}/T_{AB}$. Measurement yields

$$\frac{R_U}{R_M} \approx 2.65. \tag{6.5}$$

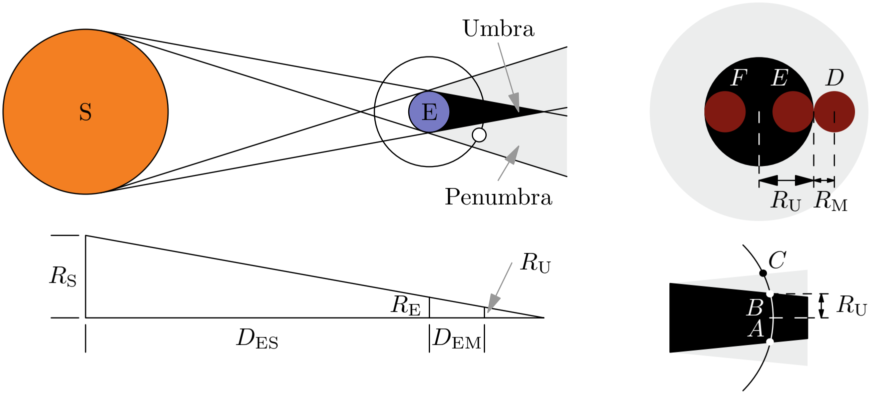


Fig. 6.3 Situation in a lunar eclipse

Aristarchus related the radius of the umbra to the other quantities. Regarding the sketch in Fig. 6.3, bottom left, one gets

$$\frac{R_S - R_E}{D_{ES}} = \frac{R_E - R_U}{D_{EM}}. \tag{6.6}$$

Although the observation of the lunar eclipse gave us one new unknown, R_U, it also gave us two new equations, (6.5) and (6.6). We now have five unknowns and five equations and can determine the unknowns.

The remainder is a matter of playing with the equations. From (6.6) and (6.1) follows

$$\frac{R_S - R_E}{R_S} = \frac{R_E - R_U}{R_M},$$

and from that,[13]

$$R_M = \frac{1 + \frac{R_M}{R_S}}{1 + \frac{R_U}{R_M}} R_E.$$

[13] The intermediate calculation steps, in case you are interested, are

$$\begin{aligned}
\frac{R_S - R_E}{R_S} &= \frac{R_E - R_U}{R_E}, \\
1 - \frac{R_E}{R_S} &= \frac{R_E}{R_M} - \frac{R_U}{R_M}, \\
1 - \frac{R_E}{R_M}\frac{R_M}{R_S} &= \frac{R_E}{R_M} - \frac{R_U}{R_M}, \\
\frac{R_E}{R_M}\left(1 + \frac{R_M}{R_S}\right) &= 1 + \frac{R_U}{R_M}.
\end{aligned}$$

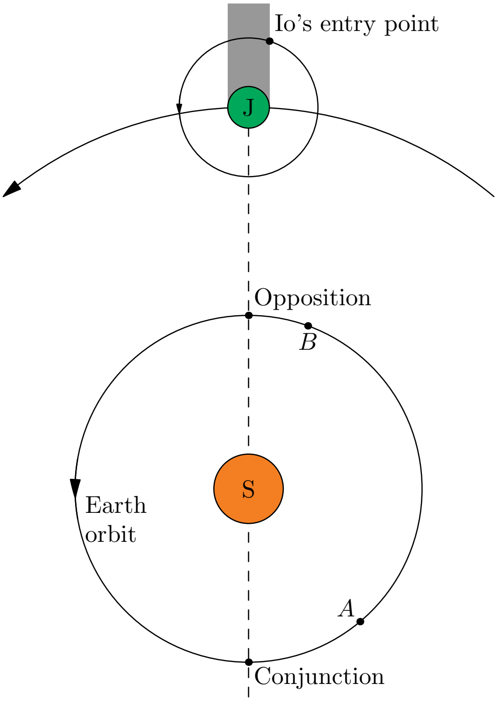


Fig. 6.4 Rømer's method for the determination of the speed of light

The quantities on the right side are known from (6.4) and (6.5). We get

$$R_{\mathrm{M}} = \frac{1 + 1/390}{1 + 2.6} R_{\mathrm{E}} = 0.27 R_{\mathrm{E}}.$$

6.3.2 Rømer's Method

In 1672, Ole Rømer performed measurements of the orbital period of Jupiter's moons.[14] The basic idea of these measurements (see Fig. 6.4) is that each moon, once per orbit, enters Jupiter's shadow and becomes eclipsed. Depending on the moon, this eclipse happens every few days and can be seen while the Earth is between the positions A and B in its orbit. The eclipse, in good approximation, happens once per moon orbit (see Exercise 17). If the orbital period of a moon is known, using multiples of it, a "timetable" with future eclipses can be compiled (for the timetable of the moon Io, see Table 6.1, column "Prediction"). Already in 1668, Cassini had found that there are systematic deviations between the predicted times and the actual

[14] Copernicus had introduced Heliocentrism in his book "De revolutionibus orbium coelestium" in 1543. So, when Rømer performed his experiment, it was already well established that the planets orbit around the Sun.

Table 6.1 Calculated timetable for Io ("Prediction") and the difference from actually observed eclipse times. Remarks: (a) conjunction, (b) entry point outshined by the Sun, (c) entry point covered by Jupiter

Transit no.	Prediction	Difference	Remarks
0	0	–	(a), (b)
1	1.769 d	–	(b)
2	3.538 d	–	(b)
...	...	...	
68	120.292 d	220 s	
69	122.061 d	220.1 s	
...	...	...	
612	1082.628 d	450 s	
...	...	...	
1150	2034.350 d	705 s	
...	...	...	
1224	2165.256 d	–	(c)
...	...	...	
2448	4330.512 d	–	(a), (c)

times (see column "Difference"). These deviations have a period of almost 400 days, which corresponds exactly to the time between two Jupiter oppositions (i.e., when the Earth lies in between the Sun and Jupiter). Rømer linked the effect to the fact that the distance between Earth and Jupiter during these 400 days changes by the diameter of the Earth's orbit, i.e., by about 300 mio. km. If the Earth were to stay at a fixed position relative to Jupiter, the prediction would be correct. The Earth, however, moves, and the distance between it and Jupiter changes. If the Earth is in position A, the light from Jupiter has to traverse more than the Earth orbit's radius, in addition to the distance when the Earth is in position B. In this way, Rømer was able to relate the duration that the light needs to traverse the diameter of the Earth's orbit to the deviations between the predicted eclipses of Jupiter's moons and the actually observed eclipses, and thus to estimate the speed of light. His goal was only to show that the speed of light is finite. For this reason, he did not optimize his method for precision. Notwithstanding, he obtained an acceptable value of 213,000 km/s.

Note that Rømer's method is very similar to the Doppler effect (see Sect. 4.2.4). Jupiter and its moon Io are the clock (which "ticks" when Io disappears in Jupiter's shadow) and the Earth is the observer, which sometimes approaches the clock and sometimes moves away from it.

Exercise 17: The time needed by the Jupiter moon Io from one eclipse to the next is a little bit larger than its orbital period. The reason is that, during Io's orbit, Jupiter moves a little bit forward on its own orbit around the Sun. Hence, the direction of the shadow changes slightly. Calculate the difference ΔT between the duration from one eclipse to the next and the moon's orbital period and compare

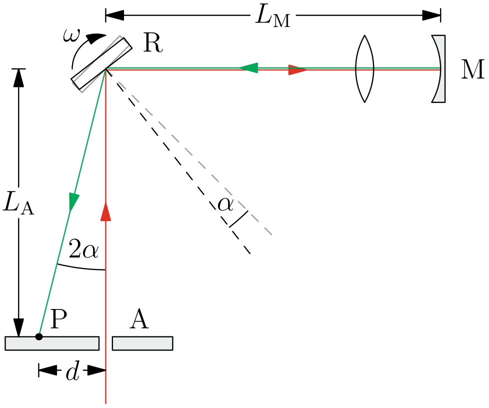


Fig. 6.5 Principle of the method of the rotating mirror. See the text

the result to the time needed by a light ray to traverse the Earth orbit's diameter. Use the following data: orbital period of Io: $T_\mathrm{Io} = 1.769$ days; orbital period of Jupiter: $T_\mathrm{Jup} = 11.86$ years.

6.3.3 *Bradley and Stellar Aberration*

In Sect. 4.4.2, we explained the physics of stellar aberration. In the course of a year, a star close to the north pole of the ecliptic geocentric coordinate system describes a small circle with a radius of $\delta = 20.5''$. Theory gives us (4.18), where $\beta_E = v_\mathrm{E}/c$ with the orbital velocity v_E of the Earth. We know the distance D_ES of the Earth and the Sun, and therefore the Earth's orbital velocity, and can determine the speed of light from $c = v_\mathrm{E}/\tan\delta$. Bradley did this in 1725 and got the very good value of 295,000 km/s for the speed of light—only 1.6% off from the modern value.

6.3.4 *The Method of the Rotating Mirror*

A measurement approach that can be performed in a laboratory is the **method of the rotating mirror**. It was proposed by François Arago in 1838, and carried out for the first time by Léon Foucault in 1850/51, and again, with a more sophisticated setup, in 1862 [Foucault62].

In this method (see Fig. 6.5), light from a lamp passes an aperture A and then hits a mirror R that rotates with a large angular velocity ω. At a particular mirror position, the reflected light traverses the distance L_M, hits another mirror M and becomes reflected back to the rotating mirror R. For the whole path from the rotating mirror R to the mirror M and back, the light needs the time $\Delta t = 2L_\mathrm{M}/c$. During this time, the rotating mirror R has been rotated by the angle $\alpha = \omega \Delta t = 2L_{\mathrm{M}}\omega/c$. After reflection, the light will no longer pass the aperture, but will hit the aperture stop next to it at a point P, the distance d from the aperture A. If the aperture has the distance L_{A} from the rotating mirror, we have $d = L_{\mathrm{A}} \tan(2\alpha)$. Then, using the approximation $\tan x \approx x$ for small x, we get $d = 4L_{\mathrm{A}}L_{\mathrm{M}}\omega/c$ and, eventually, for the speed of light,

$$c = \frac{4L_{\mathrm{A}}L_{\mathrm{M}}\omega}{d}.$$

Foucault had in his experiment $L_{\mathrm{M}} \approx 20\,\mathrm{m}$ and $L_{\mathrm{A}} \approx 1\,\mathrm{m}$. His rotating mirror was running at 400 revolutions per second ($\nu = 400\,\mathrm{Hz}$) and he determined the shift of the spot on the aperture to be $d \approx 0.7\,\mathrm{mm}$.

With the exact numbers, he got the very good value of 298,000 km/s for the speed of light. This value deviates by only 0.26% from today's value.

6.3.5 *Modern Measurement and Definition of the Value*

Definition of the speed of light. The value of the speed of light depends, of course, on the choice and definition of the units. If the second and the meter were respectively defined by a reference clock and a reference rod, the best measured value for the speed of light would be made its "official" value. Since the velocity represents the relationship between the units of time and length, with an absolute velocity, we have the opportunity to define the unit of time by the unit of length, or vice versa. Since, nowadays, times can be measured more precisely than lengths, the second alternative has been chosen, and in 1983, the meter was defined by the 17th *Conférence Générale des Poids et Mesures* (CGPM) via the second[15]:

> **Definition of the meter**: The meter is the length of the path traveled by light in vacuum during a time interval of 1/299,792,458 of a second.

With this definition, the speed of light is *defined* to be exactly $c = 299{,}792{,}458\,\mathrm{m/s}$. Usually, one uses the very good approximation of $c \approx 300{,}000\,\mathrm{km/s} = 3 \times 10^8\,\mathrm{m/s}$ in calculations. In three nanoseconds, light in vacuum covers about one meter.[16] In one second, it travels 7.5 times around the Earth or almost reaches the Moon. And in 8.5 mins., it covers the path to the Sun. Many satellites are in the geosynchronous orbit. If such a satellite is exactly above us (in the zenith), it has a distance of a little bit more than 35,000 km and light would take about 0.2 s to reach it.

[15] We will discuss the definition of the second in Sect. 9.4.

[16] The USA is exceptionally lucky here with their outdated system of units: light travels pretty much exactly one foot per nanosecond!

Note that this definition of the meter only makes sense because the principle of the absolute speed of light holds. Otherwise, the length unit would depend on the chosen reference system!

Modern measurement of the speed of light. Lasers have made it possible to determine the speed of light to a very high precision. The idea is simple. Take a monochromatic electromagnetic wave (light beam) of the form $\sin(2\pi(x/\lambda - \nu t))$ in vacuum, measure the wavelength λ and the frequency ν and calculate the speed of light from $c = \lambda\nu$. The first challenge is to find a stable wave, because the frequency (or, equivalently, the wavelength) of a laser usually changes by a minuscule amount in time. The remedy is to stabilize the laser's frequency by coupling it to a much more stable absorption line in the spectrum of an atom or molecule. It turns out that a particular laser, the helium-neon laser, and methane molecules make a perfect fit. One of the possible emission frequencies of this laser and an absorption line of methane, which can be used for the stabilization of the laser, both lie at a wavelength of 3.39 μm (this is infrared light).

Investigators around K. M. Evenson at the NBS Boulder at the beginning of the 1970s succeeded in stabilizing their helium-neon laser in this manner and measured the wavelength and the frequency with high precision [Evenson+72]. Measuring the wavelength meant comparing it to the wavelength of a particular line in the spectrum of krypton-86, which, at that time, was used to define the meter. And measuring the frequency meant comparing it to the frequency of a so-called hyperfine transition in caesium-133, which defines the second (see Sect. 9.11.2). Evenson et al. got $\nu = (88.376181627 \pm 0.000000050)$ THz for the frequency of the stabilized laser and $\lambda = (3.392231376 \pm 0.000000012)$ μm for its wavelength. Multiplying then yields a value of $c = (299{,}792{,}456.2 \pm 1.1)$ m/s for the speed of light.

Other groups of investigators made similar experiments and, eventually, the 17th CGPM, in 1983, made its decision on the definition of the meter.

Chapter 7
Relativity of Simultaneity

7.1 Introduction

After all the preparations in the last six chapters, we can finally deal with the first of the various very surprising effects of the special theory of relativity: the relativity of simultaneity. Along the way, we introduce the spacetime diagram. This diagram allows us to understand the (kinematic) effects of the special theory of relativity using only geometry and without too much mathematical hassle. In Sect. 7.3, we ask ourselves what simultaneity actually means, only to come to the point in Sect. 7.5 and realize that events that are simultaneous for Alice are not necessarily simultaneous for Bob. In Sect. 7.8, we learn why velocities faster than light are impossible (at least, in regards to the velocity of a transport of energy or information).

7.2 The Spacetime Diagram I

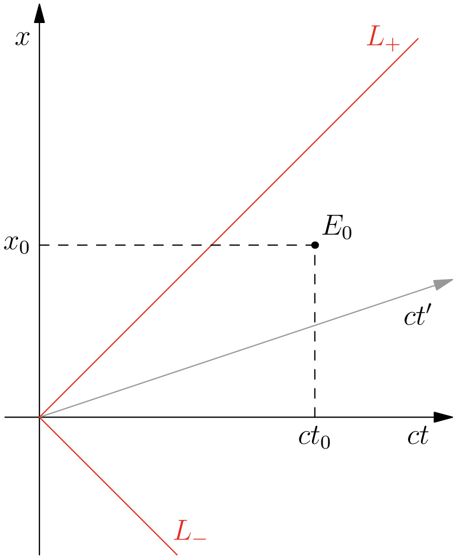
In the remainder of this book, we frequently will use diagrams that are similar to that in Fig. 7.1. They are called **spacetime diagrams** (STD) or **Minkowski diagrams** after their inventor, Hermann Minkowski. In principle, these are the x-t-diagrams known from classical mechanics.

The difference is that, in classical mechanics, the time is an absolute quantity. All observers (including Alice and Bob) have the same time. The time does not depend on the reference frame. This can easily be seen from the Galilei transformation (see Sect. 3.4.2). In classical mechanics, time is often considered as a parameter and the x-t-diagram is considered as one space dimension that is plotted against the parameter "time".

Relativity of simultaneity (which we will discuss in a jiffy) implies that, in special relativity, the time depends on the observer. Space and time in special relativity are considered on the same footing; one no longer speaks of a three-dimensional space

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_7

Fig. 7.1 Spacetime diagram



and a separate dimension of time, but rather of **four-dimensional spacetime**.[1] The spacetime diagram is a representation of spacetime. On a piece of paper, however, we can draw either one space dimension and the time dimension in a two-dimensional spacetime diagram or, two space dimensions and the time dimension in a three-dimensional spacetime diagram, in perspective drawing.

The trajectory of an object (this can be a massive object or a light pulse) in the spacetime diagram is plotted in the same way as in the x-t-diagram. A point in spacetime is called an **event**. An event $E_0 = (t_0, x_0)$ has a position x_0 (here, it is one-dimensional, but it could also be a vector in three-dimensional space) and a time t_0. The coordinates of an event (such as those of points in space) depend on the chosen coordinate system.

Frequently, one uses the time coordinate ct, the time multiplied by the speed of light. This has two advantages: (1) The time axis has the same unit as the space axes. Time is measured in meters and corresponds to the distance that light in vacuum travels in this time (30 cm correspond to about 1 ns and 300,000 m correspond to about 1 s). (2) The slope of the trajectory $x = ct$ of a light beam that travels in the positive x-direction is exactly one. When using ct as a time coordinate, the light pulse that travels through the origin of the spacetime diagram becomes the bisecting line of the coordinate axes. In Fig. 7.1, the light beam traveling in the positive x-direction is denoted by L_+ and that traveling in the negative x-direction is denoted by L_-.

Instead of the **trajectory** of an object, in the context of the spacetime diagram, one speaks of the object's **world line**. We will see later[2] that no object can move

[1] Sometimes, one talks about "3 + 1 dimensions" (or "1 + 1 dimensions") to make clear that 3 (or 1) space dimensions plus the time dimension are considered.

[2] In Sects. 7.8.1 and 13.1.3.

Fig. 7.2 The clock park in Düsseldorf's Volksgarten

faster than light. As the slope of the trajectory is the instantaneous velocity, the slope of an object's world line never can be larger than 1 (or smaller than -1).

If Bob moves relative to Alice with a constant velocity v in the x-direction and both meet in the origin of the spacetime diagram, Bob's trajectory is given by $x = vt$. At the same time, this is his time coordinate axis ct', which is defined by $x' = 0$.

7.3 Simultaneity and Synchronous Clocks

When are two events simultaneous? The concept of simultaneity is not as trivial as it seems on first sight. The naive approach does not work:

Let's say, in jest, that you observe with your telescope a bag of rice toppling over on the surface of Mars. You hold a clock in your hand. When you see the sack falling over, you read the clock. Suppose you read "one o'clock in the afternoon". In that situation, it would be wrong to say that the sack of rice fell over at one o'clock in the afternoon, because the light from Mars needed a certain amount of time to travel to the Earth. So, the sack fell over several minutes before one o'clock in the afternoon. The correction of the measured time that is necessary due to finite signal traveling times is called **retardation**. This correction must not be forgotten. But this is only the smallest problem associated with simultaneity.

Exercise 18: Determine the distance of Mars from the Earth when both are on the same ray (half-line) starting at the Sun. Calculate the traveling time of light from Mars to the Earth.

How can we demonstrate that, in a given inertial frame, two distant events are simultaneous? One could place a clock next to both events (as in the Volksgarten in Düsseldorf, see Fig. 7.2) and read the time at which an event occurs from the clock adjacent to where the event happens. If the times are the same for two events, the

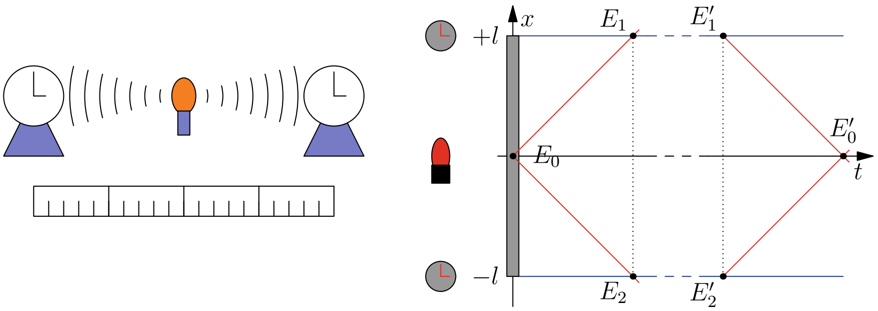


Fig. 7.3 Synchronization of clocks. Left: A lamp, two clocks and a measuring rod in space. Right: Illustration of the method in the spacetime diagram

events happen simultaneously. But then, one has to synchronize these clocks. How can that be achieved? One method is to synchronize the clocks at a common position and, after that, to place the clocks in their final positions.[3] We will use a different method (see Fig. 7.3). First, we place a lamp in the exact middle spot between the clocks. We can use a measuring rod to determine where this middle position is. Then, we switch on the lamp. Each clock will be reset to zero time when the lamp's light arrives at the clock. Both clocks will then show the same time, because the light needs the same traveling time to reach each clock. And the reasons for this are, first, the equal distances, and second, more importantly, the equal speed of light, as guaranteed by the *principle of the absolute speed of light*. Because of this principle, we know that this synchronization method works in each inertial frame.

The synchronization method is illustrated in the spacetime diagram in Fig. 7.3 on the right side. The lamp is at the location $x = 0$ and the clocks at the locations $x = \pm l$. The event E_0 stands for the switching on of the lamp and, at E_1 and E_2, the light arrives at the respective clocks. In an alternative method, the clocks at events E_1' and E_2' send a light pulse to $x = 0$. If the light pulses arrive there at the same time, the events E_1' and E_2' are simultaneous.[4]

Clock synchronization: When a light pulse in an inertial frame is sent from the location that is in the exact middle spot between two clocks and the clocks show the same time when the pulse arrives, then the clocks are synchronized.

Alternatively: Two clocks are synchronized when they send a light pulse at the same clock time in the direction of the other clock and the light pulses meet exactly in the middle between the clocks.

[3] This works, but only if the transport of the clocks is very slow. We will see the reason for this in Sect. 9.10.

[4] Note that, to define when two events **at different locations** are simultaneous, we need to know when two events **at the same location** are simultaneous.

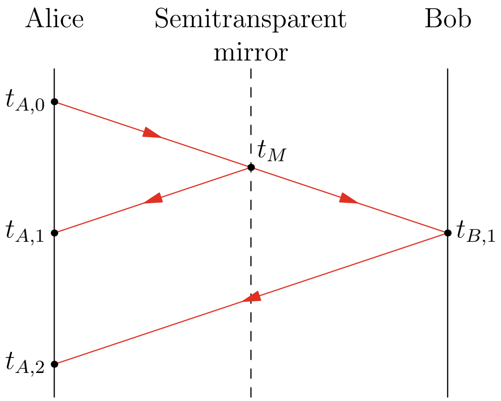


Fig. 7.4 Equivalence of Einstein synchronization and symmetric synchronization. See solution to Exercise 19

Our method works very well for determining whether two clocks are synchronous, and also for synchronizing two clocks. Sometimes, two clocks A and B have to be sychronized without changing the clock time of one of the two—say, clock A. In that case, one speaks of *synchronizing clock B according to clock A*. This is possible with our method and works as follows: the light pulse starts at the middle spot between the clocks. When it arrives at clock A, this clock's time is not changed. But the time t_A that the clock shows upon arrival of the light pulse is communicated to the other clock. When the light pulse arrives at clock B, this clock is first set to zero. Once the information from clock A arrives at clock B, the time t_A when the light pulse arrived at clock A is added to the clock time of clock B. Then, both clocks are synchronous.

It is easier, however, when clock A sends a light pulse to clock B. As soon as this pulse arrives at clock B, that clock's time is set to $t_A + D/c$, where D is the distance of the clocks (or the length of the path traveled by the light pulse).

In his original article on special relativity, Einstein gave a different method. Using **Einstein synchronization**,[5] clock A sends a light pulse to clock B at time t_0, which clock B immediately sends back to clock A. Suppose the light pulse arrives at clock A at time t_1. Then, clock B has to be adjusted such that, upon arrival of the light pulse, the clock time was $(t_1 - t_0)/2$. In inertial frames (and under the assumption that the principle of the absolute speed of light holds), all these methods are equivalent.

And again: Maybe you wonder why we used light pulses (or light signals) for the synchronization. The reason is that, due to the principle of the absolute speed of light, the synchronization methods become particularly easy. It is, however, also possible to use signals with other velocities. But this requires a different synchronization method (see Exercise 45).

[5] Actually, this synchronization method should be called *Einstein-Poincaré synchronization*, because Poincaré had already used it five years before Einstein.

Exercise 19: Show that our synchronization method with the two light pulses emitted simultaneously at the middle point between two clocks (which we will call "symmetrical synchronization") and Einstein synchronization are equivalent. To do so, place a semitransparent mirror exactly halfway between the clocks and perform an Einstein synchronization (Fig. 7.4).

7.4 Alice and Bob in Space

In the following chapters, we will conduct some *Gedanken experiments* from which we deduce important consequences of the special theory of relativity. The actors in these Gedanken experiments are Alice and Bob, representing inertial frames.[6] One would say that Alice and Bob are *inertial observers*. Each has their own coordinate system. We usually assign the "unprimed" coordinates (x, y, z) and time t to Alice and the "primed" coordinates (x', y', z') and time t' to Bob (see Fig. 7.5).

Usually, Bob will travel relative to Alice with the velocity v. We orient the two coordinate systems such that Alice's x-axis and Bob's x'-axis coincide. Furthermore, the y- and y'-axes are parallel, as are the z- and the z'-axes. Bob will move on Alice's x-axis and both meet at $t = 0$, as measured by Alice, and $t' = 0$, as measured by Bob. From Alice's perspective, Bob is located at $x = vt$, and from Bob's perspective, Alice is located at $x' = -vt'$. This constellation is called the **standard configuration**.[7]

In the special theory of relativity, Alice and Bob do not travel on trains (as they do in classical physics, see Sect. 3.4), but on spaceships in space. Alice sometimes stands on the platform of the space station and sometimes she travels with her rocket. Bob always travels with his rocket, which is of the same type as Alice's. In particular, the rockets have the same size.[8]

7.5 Simultaneity Is Relative!

The two central principles of SR, Einstein's principle of relativity and the principle of the absolute speed of light, have surprising consequences for the concept of simultaneity.

[6] If, in exceptional cases, they represent accelerated reference frames, this circumstance will be explicitly noted.

[7] This concept is taken from W. Rindler, "Relativity" (Oxford University Press, 2006), the book mentioned in the Preface.

[8] *O tempora, o mores!*: In Galileo's and Newton's time, there were, of course, neither trains nor rockets. That's why Alice was standing on the pier at that time and Bob passed her in a ship. In Einstein's times, Alice already had replaced the ship with a train and the pier with a train station. These days, Alice sits in a rocket and travels through outer space.

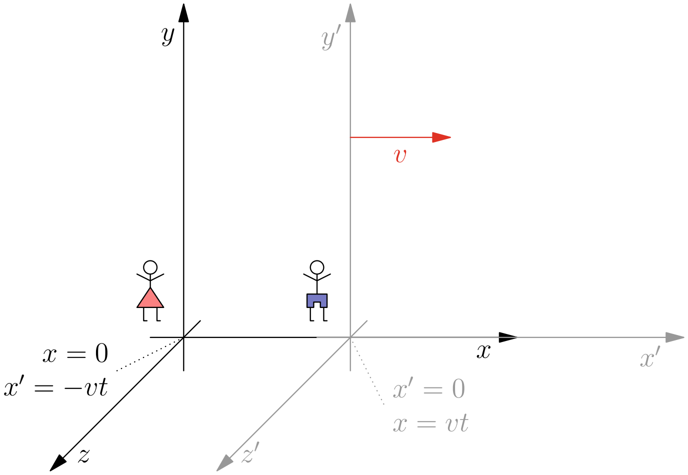


Fig. 7.5 The coordinate systems of Alice and Bob, in standard configuration

7.5.1 Gedanken Experiment

To understand these consequences, we imagine that Alice stands in a space station and Bob flies by with his rocket, having the velocity v relative to the space station (see Fig. 7.6).

Bob has fixed clocks at both ends of his rocket, the **front end** and the **rear end** (in the following, we assume that all clocks are perfectly precise). We call them the **front clock** and the **rear clock**. To synchronize these clocks as described in Sect. 7.3, Bob places himself in the center of the rocket (and, therefore, exactly midway between the clocks). Then, he switches on a lamp, also midway between the clocks, which sends light pulses in the directions of both clocks at the exact same time (see Fig. 7.6, left side). Upon arrival of the light signals, each clock is reset. Due to the fact that both clocks stand at the same distance from the lamp, the light pulses reach both clocks "simultaneously". After the procedure, the clocks run synchronously. Suppose now that Bob switched on the lamp at the exact moment that he passed by Alice.

How does Alice interpret the synchronization procedure? From her point of view, Bob is sitting in the center of his rocket (and the lamp is exactly midway between the clocks). She observes how Bob switches on his lamp and sees the light pulses (see Fig. 7.6, right side). Alice now measures the velocity of both light pulses emitted by Bob's lamp. For both, she gets c, the speed of light, exactly according to the principle of the absolute value of the speed of light. Due to the principle of relativity, this principle is valid in the same way for both Alice and Bob.

From Alice's point of view, Bob's rear clock moves toward her while the front clock moves away from her. For this reason, and again *from Alice's point of view*, the light pulse emitted by the lamp arrives at Bob's rear clock *before* it arrives at Bob's front clock. From Alice's point of view, Bob's clocks are not simultaneously

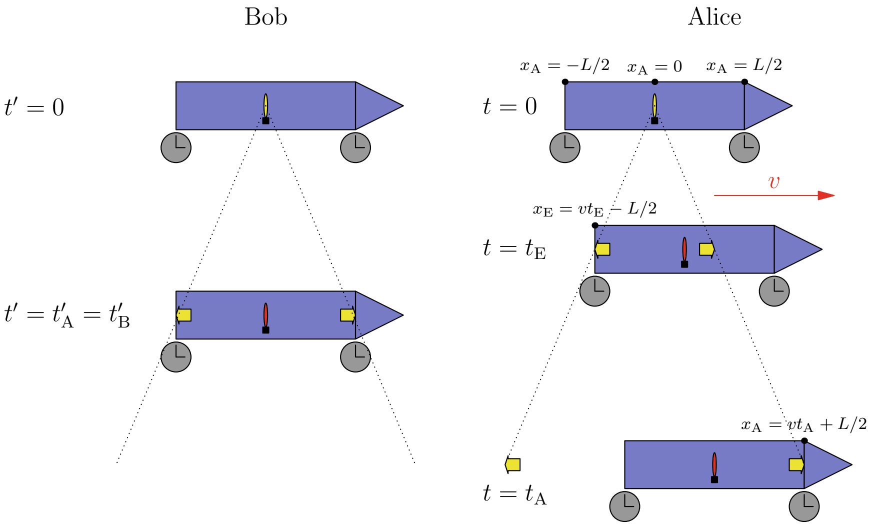


Fig. 7.6 The concept of simultaneity. Left: As seen from Bob's rocket. Right: As seen from Alice at the space station

reset. *Simultaneity for Bob is not equivalent to simultaneity for Alice*. This is exactly what the **relativity of simultaneity** means. The relativity of simultaneity is a fundamental insight of SR and strongly contradicts the concept of space and time in classical mechanics. Simultaneity is no longer absolute, it is relative! Events that are simultaneous in one inertial frame are not so in other inertial frames (in general). It makes no sense to speak about simultaneity without specifying the inertial frame for which it holds, in the same way as it makes no sense to talk about a velocity without specifying relative to which reference system it holds.

Relativity of simultaneity is **the central effect** of special relativity. Make sure that you understand the effect and, in particular, why it necessarily follows from the two principles of SR. All other (kinematic) effects of SR build upon the relativity of simultaneity.

Relativity of simultaneity: The simultaneity of two events is relative. Events that are simultaneous in a given inertial frame, in general, are not simultaneous in another inertial frame.

One could speculate that simultaneity, as defined by our synchronization method, is only relative because our synchronization procedure is deficient. Is there another synchronization procedure that yields an absolute simultaneity? We will discuss this and related questions in Sect. 7.7.

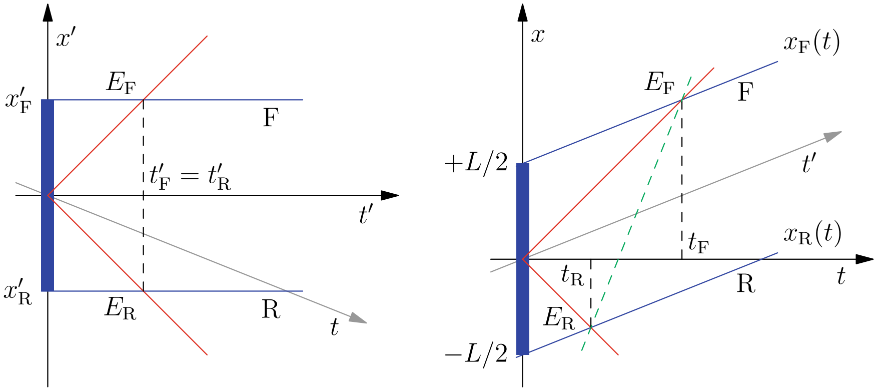


Fig. 7.7 The concept of simultaneity, represented in the spacetime diagram. Left: Bob's point of view. Right: Alice's point of view

7.5.2 The Difference from Classical Physics

How would the discussion of the Gedanken experiments in Fig. 7.6 differ in classical physics? When measuring the speed of the light pulses in Bob's rocket, Alice would not get the speed of light. Instead, the light pulse traveling to the rear clock would have the velocity $c - v$, while the light pulse traveling to the front clock would have the velocity $c + v$. This difference in velocities would specifically cancel the fact that, for Alice, the light pulses have to travel different distances before arriving at the clocks. Also, Alice would find that the light pulses arrive at both clocks at the same time (see Exercise 45).

This synchronization procedure, however, is not the same in all inertial frames. It breaks Einstein's principle of relativity, and thus we cannot use it. In classical physics, instead of the light pulse, we could use a signal that is infinitely fast. Only the special theory of relativity, however, has shown that such a signal cannot exist.

7.5.3 Spacetime Diagram

We now discuss the situation using a spacetime diagram. First, we take Bob's point of view (see Fig. 7.7, left side). His coordinates are (t', x'). The center of the rocket is located at $x' = 0$. F and R are the respective world lines (trajectories) of the front and rear ends of the rocket. At time $t' = 0$, Bob, at location $x' = 0$ sends a light pulse in both directions of the x'-axis. At event E_F, the light pulse arrives at the front end of the rocket, and at event E_R, at its rear end. Both events, by definition, are simultaneous for Bob and the line between the events is parallel to the x'-axis.

Next, we take Alice's point of view (see Fig. 7.7, right side). Suppose that Bob's rocket, *for Alice*, has the length L. (In Chap. 8, we will see that the lengths of objects at large velocities depend on the reference system. Therefore, we have to specify *for whom* this length holds. This, however, is not an issue here.) The front end and the rear end of Bob's rocket, for Alice, have the respective coordinates $x_{\mathrm{F}}(t) = L/2 + vt$ and $x_{\mathrm{R}}(t) = -L/2 + vt$ (see Fig. 7.7). At time $t = 0$, Bob's lamp emits the light pulse at $x = 0$. According to the principle of the absolute speed of light, for Alice, the light pulses also travel with velocity c in both directions of the rocket and arrive at the events E_{F} and E_{R} at the ends of the rocket.

For Alice, the forward-directed light pulse moves according to $x = ct$ and the backwards-directed one according to $x = -ct$. Because the back of the rocket moves toward the light pulse, the backwards-directed pulse first arrives at the rear clock. Call this arrival time t_{E}, then, from $-ct_{\mathrm{E}} = -L/2 + vt_{\mathrm{E}}$ follows $t_{\mathrm{E}} = (L/2)/(c + v)$. A little bit later, the forward-directed light pulse arrives at the front clock, and the arrival time is $t_{\mathrm{A}} = (L/2)/(c - v)$. It is immediately clear that $t_{\mathrm{A}} \neq t_{\mathrm{E}}$. The two events "light pulse arrives at rear clock" and "light pulse arrives at front clock" are simultaneous for Bob (according to our definition of simultaneity), but not for Alice!

With the relativity of simultaneity, *space and time became much more similar* than they are in classical physics. The following observation already holds in classical physics, and we consider it to be trivial:

> Events that, for Alice, happen *at the same location* (meaning that they have the same coordinate values) do not, in general, do the same for Bob.

Consider a car being driven. The car emits a banging noise, and then stops. For the driver, the bang and the car stopping happen at the same location, namely, in the car. For the observer at the roadside, the bang does not happen where the car stops. We could speak of the *relativity of equilocality*.

In special relativity, there is the following additional observation:

> Events that, for Alice, happen *at the same time* do not, in general, do the same for Bob.

This is an exact statement of the *relativity of simultaneity*. Space and time appear in an analogous way in both statements above.

7.6 The Spacetime Diagram II: Simultaneity

We again have a look at the spacetime diagram in Fig. 7.7, right side. There, we have drawn the x-axis of Alice. All events on this axis are simultaneous for Alice (i.e., for Alice, they happen at the same time), because for all events E on the x-axis, we have $t_E = 0$. For events that are on a parallel line to the x-axis, the same is true: all of these events happen simultaneously for Alice. Now, let us look at the t-axis. All events on this axis happen at the same position for Alice, namely, $x = 0$. The same applies to a parallel line.

What about Bob's coordinate system? For Alice, Bob travels on the line $x = ct$, and all events on this line happen at the same place for Bob, namely, his own position. So, this line is Bob's time axis, the t'-axis. What about **Bob's "axis of simultaneity"**,

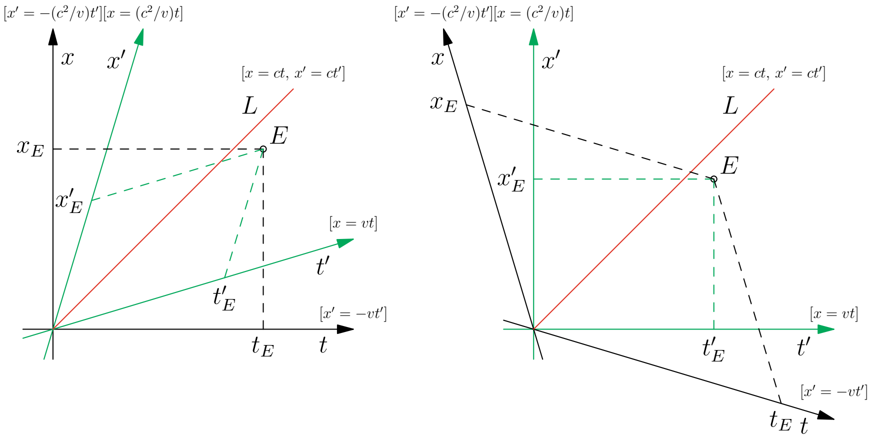


Fig. 7.8 Left: construction of the axes of Bob's coordinate system (green), as seen by Alice (black). The red line L represents a light pulse emitted by Alice in the positive x-direction. Right: the same, as seen by Bob

the x'-axis? In classical physics, Bob's x'-axis and Alice's x-axis would coincide, because, there, simultaneity is absolute (independent of the inertial frame). In special relativity, this is not true. As we learned in the last chapter, the events E_F and E_R in Fig. 7.7, right side, are simultaneous for Bob. His x'-axis must be parallel to the line through these two events, the green line in the figure (see Exercise 20). In Fig. 7.8, left side, we have drawn Bob's x'-axis. One recognizes immediately that the light pulse L bisects the angle between the x'- and the t'-axis. This must be the case, because, for Bob, the world line of the light pulse must be given by $x' = ct'$.

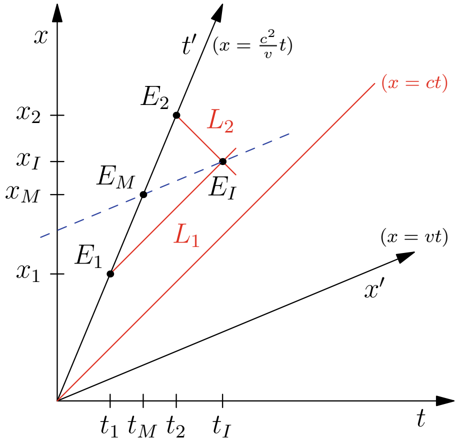
Note how the coordinates for a particular event are read from the coordinate axes. The event E in Alice's coordinate system has the coordinates (t_E, x_E), while in Bob's coordinate system, it has the coordinate (t'_E, x'_E). To read the latter, one has to draw lines that are parallel to the coordinate axes and that pass through the event E.

Bob would draw his coordinate axes x' and t' perpendicular to each other (see Fig. 7.8, right side). For Bob, Alice travels on the world line $x' = -vt'$. Her "axis of simultaneity", the x-axis, must be such that the trajectory L of the light pulse is again the angle bisector of the t- and the x-axis.

Now we know how to draw Bob's coordinate axes into Alice's spacetime diagram. What is missing still is the **scale** of Bob's axes. It would be wrong to transfer the scale of Alice's axes with a compass to Bob's axes. The reason for this is the fact that, while space may be Euclidean, this is not the case for spacetime.[9] We will see later what this exactly means. For the moment, it is sufficient to memorize the fact that angles and lengths in the spacetime diagram must not be measured with a rod!

[9] If the term "space" occurs in the expression "Euclidean space" or something similar then "space" is meant in the mathematical sense. Otherwise, "space" refers to the three-dimensional physical space.

Fig. 7.9 To Exercise 20



Exercise 20: Show that, from Alice's perspective, all events on Bob's x'-axis are simultaneous for Bob.

To do so, take two arbitrary events E_1 and E_2 on the x'-axis that, for Bob, have the coordinates $E_1 = (0, x'_1)$ and $E_2 = (0, x'_2)$ and, for Alice, have the coordinates $E_1 = (t_1, x_1)$ and $E_2 = (t_2, x_2)$, respectively. Show that light pulses emitted in these events and toward each other meet exactly in the middle: $x'_\text{M} = (x'_2 + x'_1)/2$ (see Fig. 7.9).

1. Express the coordinates of the events E_1 and E_2 as functions of t_1 and t_2.
2. Calculate Alice's coordinates of the intersection event E_I of the light lines through E_1 and E_2 (i.e., x_I and t_I as functions of t_1 and t_2).
3. Determine the line through E_I, which is parallel to Bob's time axis, and then Alice's coordinates (t_M, x_M) of the intersection E_M of this line with Bob's space axis. Express t_M as a function of t_1 and t_2.
4. If you made no mistake, you got $t_\text{M} = (t_1 + t_2)/2$ and $x_\text{M} = (x_1 + x_2)/2$, so the event E_M lies exactly midway between E_1 and E_2.

Therefore, the events E_1 and E_2 are simultaneous for Bob, because E_I happens exactly midway between them. These arguments hold for all event pairs E_1 and E_2 on the line $x = (c^2/v)t$, which, for this reason, is a space axis for Bob.

Exercise 21: Draw a spacetime diagram with Alice and Bob. Now, Bob moves in the negative x-direction. Keep the convention that Alice's coordinate axes are perpendicular to each other.

7.7 Digression: Further Thoughts on Simultaneity

7.7.1 Introduction

Conventionality of simultaneity. Einstein synchronization looks somehow arbitrary. Is it the only method for synchronization? We show that, indeed, using it is *convention*, and that there are other possible methods for clock synchronization. These alternatives, however, do not lead to different physics, but they do lead to a different notion of simultaneity. This is what is meant by the term "conventionality of simultaneity". What makes these alternatives less interesting is that they produce theories that are much more complicated than special relativity.

One-way and two-way speed of light. Let us start our digression with a look at the principle of the absolute speed of light. It states that the speed of light is independent of the direction. To be able to make practical use of this principle, we have to be able to determine the velocity of a light pulse that travels from a location A to another one B (which are all at rest relative to the observer). This requires measuring the distance ℓ_{AB} between A and B and the time Δt_{AB} that the light pulse needs to travel this distance. For the latter, one needs a clock at A and another one at B, and these clocks *must be synchronized*. Then, $c = \ell_{AB}/\Delta t_{AB}$. This velocity is called the **one-way speed of light**.

Without a synchronization method, the concept of the one-way speed of light is meaningless, and with it, the principle of the absolute speed of light, which refers to the one-way speed of light, is pointless.[10]

Indeed, determining *any velocity* requires a method for clock synchronization, which is convention. Different methods for clock synchronization yield different one-way velocities. Recall **Einstein's synchronization method**: there are two locations A and B, each of which has a clock and each of which is at rest relative to the other and the observer. A light pulse is sent from A to B and immediately back to A. The light pulse is emitted at t_1 (as measured by the clock at A), arrives at B at t_2 (as measured by the clock at B), where it is reflected and eventually arrives back at A at t_3 (as again measured by the clock at A). Then, Einstein synchronization sets $t_2 = t_1 + (t_3 - t_1)/2$.

[10] Therefore, before introducing the principle of the absolute speed of light in Sect. 6.1, from a logical point of view, we should have presented Einstein's synchronization method. The synchronization method without the principle of the absolute speed of light is useful, while the principle without the synchronization method is not. On the other hand, the principle suggests a synchronization method: At $t_A = 0$, send a light pulse to B, and when it arrives, set the time of clock B to $t_B = \ell/c$. Indeed, Einstein in his epoch-making paper on special relativity, first defined simultaneity [Einstein05a]: "Die letztere Zeit [gemeinsame „Zeit"] kann nun definiert werden, indem man *durch Definition* festsetzt, dass die „Zeit", welche das Licht braucht, um von A nach B zu gelangen, gleich ist der „Zeit", welche es braucht, um von B nach A zu gelangen." and in English: "The latter time [common "time"] can now be defined by stating *by definition* that the "time" which the light needs to get from A to B is equal to the "time" it takes to get from B to A."

Einstein's method for clock synchronization *implies that the one-way speed of light from A to B is equal to that from B to A*. Let us imagine that we are located at A and send the light pulse to B, where is gets reflected back to A, and use c_+ for the one-way speed of light from A to B ("forward one-way speed of light") and c_- for the one-way speed of light from B to A ("backward one-way speed of light"). Then, Einstein synchronization makes the forward and backward one-way velocities equal:

$$\begin{aligned} c_+ &:= \frac{\ell_{AB}}{t_2 - t_1} = \frac{\ell_{AB}}{t_1 + (t_3 - t_1)/2 - t_1} = \frac{2\ell_{AB}}{t_3 - t_1}, \\ c_- &:= \frac{\ell_{AB}}{t_3 - t_2} = \frac{\ell_{AB}}{t_3 - t_1 - (t_3 - t_1)/2} = \frac{2\ell_{AB}}{t_3 - t_1}. \end{aligned}$$

Now, if we send a light pulse from A to B and back to A to measure the **average speed** along this path, we do not need a clock at B and there's no need to synchronize. The clock at A is sufficient. This average of the velocity from A to B and that back from B to A is called the **two-way speed of light**. The two-way speed of light can be determined without a synchronization method. It is not convention.

Experiment, or: the objective physical world. One can show that there is no way to measure a (one-way) velocity without synchronizing clocks, i. e., choosing a particular synchronization method.[11] And the value of the (one-way) velocity will depend on this convention, i. e., the chosen synchronization method. One can also show that any experiment in which the light follows a closed path, such as the Michelson-Morley experiment, measures the two-way speed of light.

The Michelson-Morley experiment, in combination with two other experiments (the Ives-Stilwell and Kennedy-Thorndike experiments, see Sects. 9.7 and 9.8), shows that the two-way speed of light is independent of direction and inertial frame (see Sect. 11.4). Therefore, without choosing a synchronization method, one can show that *the two-way speed of light is independent of the direction and the inertial frame*.

The principle of the absolute speed of light in special relativity refers to the one-way speed of light that follows from the constancy of the two-way speed of light plus the Einstein synchronization. The fact that we cannot measure the one-way speed of light without a convention shows that this principle is stronger than needed. We will show that weakening the principle of the absolute speed of light to turn into the principle of the absolute *two-way* speed of light allows for a whole family of alternative theories with different synchronization methods. Lorentz's ether theory is among these theories. But all of these theories predict the same physics as special relativity. Of these theories, special relativity by far is the most simple and most elegant one. Therefore, people prefer this theory to others.

[11] Although some physicists are still arguing, and propositions to measure one-way velocities sometimes appear, this seems to be the consensus in the community. But if, indeed, we could measure one-way velocities, this would imply that only a particular one of the synchronization methods would be correct. However, for the effects predicted by special relativity, this would not change anything.

Although the one-way speed of light cannot be measured (without the convention), one can show that the one-way speed of light is independent of the velocity of the light source. Brecher did this, as we explained in Sect. 5.3.3. Suppose there's a moving source and a source at rest at A. Both emit light at the same time. At B, we find that these light pulses arrive at the same time. Here, we don't need synchronized clocks.

Alternative synchronization methods. What are these different synchronization methods that are different from Einstein's method?

Reichenbach [Reichenbach58] introduced the whole family of different synchronization methods, and Einstein synchronization is one of these. Synchronization methods different from Einstein's not only yield direction-dependent *one-way* speeds of light, but also imply different Lorentz transformations, the *generalized Lorentz transformations* [Anderson+98]. The relativistic effects like length contraction, time dilation, and so on, however, also result from these generalized transformations, which we will derive in a minute.

Reichenbach synchronization. The generalization of Einstein's synchronization method is the *Reichenbach synchronization*. Reichenbach uses the same experiment as Einstein, but instead of $t_2 = t_1 + (t_3 - t_1)/2$, he sets

$$t_2 = t_1 + \epsilon \cdot (t_3 - t_1) \tag{7.1}$$

for a particular fixed ϵ. The synchronization parameter ϵ must obey $0 < \epsilon < 1$, otherwise signals would arrive before they have been sent, which contradicts causality. The choice of $\epsilon = 1/2$ for all inertial frames amounts to Einstein synchronization.[12]

With this synchronization method, *the one-way speed of light depends on the direction*. Suppose that ℓ_{AB} is the distance between A and B. Then, the speed of light c_+ from A to B becomes

$$c_+ = \frac{\ell_{AB}}{t_2 - t_1} = \frac{\ell_{AB}}{\epsilon \cdot (t_3 - t_1)} = \frac{c}{2\epsilon},$$

where $c := 2\ell_{AB}/(t_3 - t_1)$ is the two-way speed of light. In the same way, one gets

$$c_- = \frac{c}{2(1-\epsilon)}$$

for the speed of light from B to A.[13] Furthermore, we have

$$\frac{2}{c} = \frac{1}{c_+} + \frac{1}{c_-},$$

which says that c is the harmonic mean of c_+ and c_-.

[12] In Reichenbach synchronization, each inertial frame could have its own ϵ.

[13] Note that, for $\epsilon = 0$ or $\epsilon = 1$, one of the two velocities $c_\pm$ would become infinity.

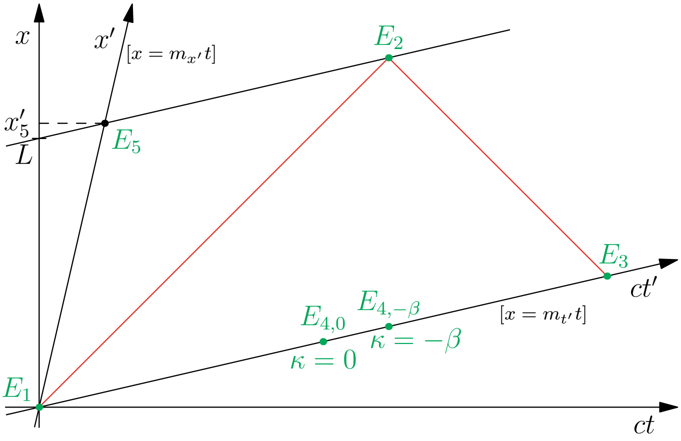


Fig. 7.10 For the construction of the generalized Lorentz transformation

7.7.2 *Generalized Lorentz Transformation*

We derive now the generalized Lorentz transformations, which are induced by the different synchronization methods.

Synchronization methods and the axis of simultaneity. The main physical ideas can be seen in 1+1 dimensions, therefore, we restrict ourselves to this situation. Then, we introduce Alice's reference frame Σ with coordinates (t, x), in which we assume that the speed of light is isotropic (independent of the direction). This implies that Alice uses Einstein's synchronization method to synchronize clocks.[14] Then, we have a second reference frame Σ with coordinates (t', x'), which is Bob's and the origin of which moves according to $x = vt$ in Σ'.

To synchronize clocks, at event E_1, Bob sends a light signal to a mirror, where it is reflected in E_2 and arrives back at Bob in E_3 (see Fig. 7.10). The mirror's world line is $x = L + vt$. Next to the mirror is a clock of the same type as Bob's clock, and the task now is to synchronize this clock. To synchronize it means to relate its time to one of the points on the ct'-axis between E_1 and E_3 and, in the moment when the light signal is reflected, set it to this time. We can use the parameter ϵ introduced by Reichenbach and write the time to which the clock at the mirror is set, as $t_2 = t_1 + \epsilon \cdot (t_3 - t_1)$, but we prefer to use the alternative parameter κ (possibly introduced in [Anderson+98]), where

$$t_2 = \frac{1-\kappa}{2} t_1 + \frac{1+\kappa}{2} t_3,$$

[14] Remember that the isotropy of the speed of light and the Einstein synchronization method are intertwined.

and which interpolates linearly from $t_2 = t_1$ for $\kappa = -1$ to $t_2 = t_3$ for $\kappa = +1$. Einstein synchronization corresponds to the case in which $\kappa = 0$. For reasons of causality, we restrict κ to $-1 < \kappa < +1$. The relation to ϵ is given by $\epsilon = (1+\kappa)/2$ or $\kappa = 2\epsilon - 1$.

For each $-1 < \kappa < +1$, we have a different synchronization method. Different synchronization methods imply different notions of simultaneity and a different x'-axis. The next step that we perform is to determine the x'-axis, for a given κ.

To determine the slope of the x'-axis as Alice sees it, we have to determine Alice's coordinates of the events E_1, E_2, E_3 (in Fig. 7.10). For E_1, which is the origin, this is trivial. For E_2, we have to determine the intersection of the mirror's world line with the light cone. This gives us

$$E_2 : (t_2, x_2) = \left(\frac{L/c}{1-\beta}, \frac{L}{1-\beta} \right).$$

Event E_3 is the intersection of the light cone starting at E_2 and traveling in the negative x-direction with Bob's world line, i. e., of $x = -c(t - t_2) + x_2$ with $x = vt$. This yields

$$E_3 : (t_3, x_3) = (2L\gamma^2/c, 2L\gamma^2\beta).$$

The event $E_{4,\kappa}$, which depends on κ, is defined by the synchonization method and given by

$$E_{4,\kappa} : (t_{4,\kappa}, x_{4,\kappa}) = \left(\frac{1+\kappa}{2} t_3, \frac{1+\kappa}{2} x_3 \right) = (L(1+\kappa)\gamma^2/c, L(1+\kappa)\gamma^2\beta).$$

The slope $m_{x'}$ of the x'-axis is then

$$m_{x'} = \frac{x_2 - x_{4,\kappa}}{t_2 - t_{4,\kappa}} = c \frac{\frac{1}{1-\beta} - (\kappa+1)\gamma^2\beta}{\frac{1}{1-\beta} - (\kappa+1)\gamma^2} = c \frac{1+\kappa\beta}{\kappa+\beta}.$$

For *Einstein synchronization* ($\kappa = 0$), we get $m_{x'} = c/\beta = c^2/v$, a fact that we already know.

But an **absolute simultaneity** is also possible. If we choose $t_{4,\kappa} = t_2$, then the x'-axis is parallel to the x-axis and events that, for Alice, are simultaneous are also simultaneous for Bob. The condition $t_{4,\kappa} = t_2$ is satisfied for $\kappa = -\beta$. We refer to the associated synchronization method as **absolute synchronization**.

From a physical point of view, there is a large difference between Einstein's synchronization method and absolute synchronization. Einstein synchronization is an *internal synchronization method*. To carry it out, only the rest frame of the observer is relevant and no reference to any other inertial frame is needed. Absolute synchronization, however, singles out one inertial frame, the aether frame, and synchronizes all other inertial frames in such a way that there is absolute simultaneity. This can only be performed if the aether frame is known. Absolute synchronization is an *external synchronization method* and the parameter β in absolute synchronization is the velocity of the inertial frame *relative to the aether*.

So, from all the different synchonisation methods for $-1 < \kappa < 1$, the two that interest us most are Einstein synchronization with $\kappa = 0$ and absolute synchronization with $\kappa = -\beta$.

Generalized Lorentz transformation. Now, we look for the most general linear transformation, which maps the t-axis to the t'-axis and the x-axis to the x'-axis. We know already that the t'-axis, for Alice, is given by $x = vt$ and the x'-axis by $x = m_{x'}t$. If we write the general linear transformation in the form

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} a & b \\ d & e \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix}$$

with parameters a, b, c, d that may depend on κ and β, these conditions yield

$$\begin{aligned} ct' &= act + bx = 0 \quad \curvearrowright \quad x = -\frac{a}{b}ct \quad \curvearrowright \quad -\frac{a}{b} = m_{x'}, \\ x' &= dct + ex = 0 \quad \curvearrowright \quad x = -\frac{d}{e}ct \quad \curvearrowright \quad -\frac{d}{e} = \beta \end{aligned}$$

and the following form for the general Lorentz transformation:

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \hat{T}_\kappa \begin{pmatrix} ct \\ x \end{pmatrix} \quad \text{with} \quad \hat{T}_\kappa = \begin{pmatrix} a & -a\frac{\kappa+\beta}{1+\kappa\beta} \\ -e\beta & e \end{pmatrix},$$

where a and e still have to be determined.

Two-way speed of light. We determine the parameters a and e by requiring that the two-way speed of light be independent of the reference frame. We know that the two-way velocity is not convention, and the Michelson-Morley experiment has demonstrated that, indeed, it is independent of the inertial frame.

For Bob, this means that the time t'_3 that the light needs to travel to the mirror and back must be equal to the total distance traveled (to-and-fro) divided by c:

$$t'_3 \overset{!}{=} \frac{2L'}{c}.$$

We start calculating t'_3, proceeding as follows:

$$\begin{aligned} ct'_3 &= a\left(ct_3 + \frac{\kappa+\beta}{1+\kappa\beta}x_3\right) = a\left(1 - \beta\frac{\kappa+\beta}{1+\kappa\beta}\right)ct_3 \\ &= a\gamma^{-2}\frac{1}{1+\kappa\beta}ct_3 = a\gamma^{-2}\frac{1}{1+\kappa\beta}2L\gamma^2 \\ &= 2aL\frac{1}{1+\kappa\beta}. \end{aligned}$$

To determine L', which is equal to Bob's x'-coordinate x'_5 of the event E_5, we first calculate ct_5 from the intersection of the mirror's world line with the x'-axis

$$L + \beta ct_5 = \frac{1+\kappa\beta}{\kappa+\beta} ct_5,$$

from which follows

$$ct_5 = \gamma^2 \cdot (\kappa + \beta) L$$

and

$$x_5 = \frac{1+\kappa\beta}{\kappa+\beta} ct_5 = \gamma^2 \cdot (\kappa+\beta) L,$$

and eventually, by applying the transformation $\hat{T}_\kappa$,

$$L' = x'_5 = e \cdot (x_5 - \beta ct_5) = e\gamma^2 L((1+\kappa\beta) - \beta(\kappa+\beta)) = eL.$$

With all these preparations, from $ct'_3 = 2L'$, we finally get

$$2aL \frac{1}{1+\kappa\beta} = 2eL \quad \text{or} \quad \frac{a}{e} = 1 + \kappa\beta.$$

Therefore, the generalized Lorentz transformation becomes

$$\hat{T}_\kappa = e \begin{pmatrix} 1+\kappa\beta & -(\kappa+\beta) \\ -\beta & 1 \end{pmatrix}$$

and only the parameter e is left. To determine it, consider Bob's clock at $x = vt$. For this clock, $ct' = e((1+\kappa\beta)ct - (\kappa+\beta)\beta ct) = (e/\gamma^2)ct$. This describes time dilation, and from the experiment (e. g., Ives-Stilwell experiment), we know that $\Delta t' = \gamma^{-1}\Delta t$. Therefore, we have $e = \gamma$ and

$$\hat{T}_\kappa = \gamma \begin{pmatrix} 1+\kappa\beta & -(\kappa+\beta) \\ -\beta & 1 \end{pmatrix} \tag{7.2}$$

as the final form of the **generalized Lorentz transformation**.

For Einstein synchronization ($\kappa = 0$), this becomes

$$\hat{L} = \hat{T}_0 = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix},$$

which is the Lorentz transformation $\hat{L}$, and for absolute synchronization ($\kappa = -\beta$), we get

$$\hat{T}_{\text{as}} := \hat{T}_{-\beta} = \begin{pmatrix} 1/\gamma & 0 \\ -\gamma\beta & \gamma \end{pmatrix},$$

or $t' = \gamma^{-1}t$, which clearly leads to absolute simultaneity.

Independently of κ, and, in particular, for the two considered cases, we get the same results for time dilation and length contraction, as well as the derived effects like the relativistic Doppler effect, the correct formula for aberration, the twin paradox, and so on.

"Time shift". Indeed, all $\hat{T}_\kappa$ for the same β are very similar. This can be seen if we write a general Lorentz transformation as a Lorentz transformation and a further transformation $\hat{S}_\kappa$, i. e., $\hat{T}_\kappa = \hat{S}_\kappa \hat{T}_0$. Then, $\hat{S}_\kappa$ is given by

$$\hat{S}_\kappa = \hat{T}_\kappa \hat{T}_0^{-1} = \gamma \begin{pmatrix} 1+\kappa\beta & -(\kappa+\beta) \\ -\beta & 1 \end{pmatrix} \gamma \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix} = \begin{pmatrix} 1 & -\kappa \\ 0 & 1 \end{pmatrix}.$$

In other words: the general Lorentz transformation $(ct, x) \to (ct'', x'')$ can be seen as first conducting a Lorentz transformation $(ct, x) \to (ct', x')$ and then the transformation

$$\begin{aligned} ct'' &= ct' - \kappa x' \\ x'' &= x', \end{aligned} \tag{7.3}$$

which is nothing but a shift of the time scale by an amount that is dependent on x'.

7.7.3 *Lorentz's Ether Theory Versus Einstein's Special Relativity*

Based on the idea of a *motionless luminiferous aether* (which is nothing but a special inertial frame) and the validity of Maxwell's electrodynamics, Lorentz, in 1892, began to develop a theory of electrodynamics for inertial frames that move relative to the aether. He called this the "theory of electrons", and later on, it became **Lorentz's ether theory** (LET). Very consistent with the aether, this theory is based on an **absolute time** (or *true time*), which, in all inertial frames, is the same as in the aether system. This corresponds to what we called absolute synchronization. Lorentz, in addition to the absolute time, in 1892, already had introduced the **local time** (exactly by (7.3)). Lorentz needed local time to explain the findings in experiments like stellar aberration and Fizeau's, as well as the Doppler effect, but he never believed it was the true time. Poincaré later noticed that local time results from what today we call Einstein synchronization (and therefore, we often speak about Poincaré-Einstein synchronization) and is the time concept used in special relativity. To be able also to explain the result of the Michelson-Morley experiment, Lorentz had to introduce the Lorentz-FitzGerald contraction into his theory. Based on the groundwork of others, Lorentz then derived what Poincaré, in 1905, called the **Lorentz transformations** (Lorentz did not yet have the prefactor γ; this was accomplished by Poincaré). In that same year, Poincaré also showed that Maxwell's equations were fully **form-invariant** when subjected to a Lorentz transformation. He further demonstrated that the "mechanisms" of Lorentz's ether theory (local time with time dilation and length contraction) conspired in a way that it would never be possible to detect the aether: Lorentz's ether theory builds upon the fundament of the luminiferous aether, but then proves that this aether never can be detected.

In the end, Lorentz's ether theory makes the same physical predictions as special relativity. The key differences are:

- The fact that the *principle of relativity* does not hold. The reason for this is the assumption of the luminiferous aether. The theory, however, is form-invariant with respect to Lorentz transformations, and it actually does not do any harm to remove the aether. With this modification, Lorentz's ether theory obeys the relativity principle.
- The focus on *absolute time*. The needed ingredients, including the *local time*, have been part of Lorentz's ether theory from the beginning on. Putting the local time instead of the absolute time at center stage makes Lorentz's ether theory equal to special relativity in this aspect. Absolute time, by its definition, is tied to the existence of the aether.
- The *law for velocity addition* is not symmetric in Lorentz's ether theory (not even in one dimension). This is due to the fact that velocities are defined on the basis of absolute time, which renders the formula more complicated, as in special relativity. The Lorentzian addition of velocities in special relativity is a lot simpler and based on the definition of velocities with what is the local time in Lorentz's ether theory (it is funny that the velocity addition formula in special relativity carries Lorentz's name, despite the fact that, in his own theory, the velocity addition formula is different).
- There is *no unification of space and time* in Lorentz's theory. This unification is only possible if one takes Lorentz's local time seriously.

To sharpen these findings: on the basis of the existence of the luminiferous aether, Lorentz's ether theory comes to the conclusion that the aether is never observable. Just by removing the aether from this theory and performing the small necessary conceptual changes in the edifice, it becomes equal to Einstein's special relativity.

This concludes our digression on the conventionality of simultaneity.[15]

[15] Most of this wisdom comes from the three papers by Mansouri and Sexl [MansouriSexl77a, MansouriSexl77b, MansouriSexl77c], who performed a similar derivation of the Lorentz transformation from experimental findings as Robertson [Robertson49]. Robertson, however, took Einstein's synchronization method for granted, and Mansouri and Sexl also considered other synchronization schemes. The conventionality of simultaneity was pointed out earlier by Reichenbach [Reichenbach58]. Anderson et al. [Anderson+98] extended Mansouri and Sexl's analysis considerably, illuminated the effects on physics and contributed strongly to a convergence of different opinions. Selleri [Selleri94] then formulated the derivation that we followed here. Rizzi et al. [Rizzi+08] disputed and proved wrong several of the physical implications stated by Selleri (but not his derivation). A nice and readable summary was written by de Abreu and Guerra [deAbreuGuerra15].

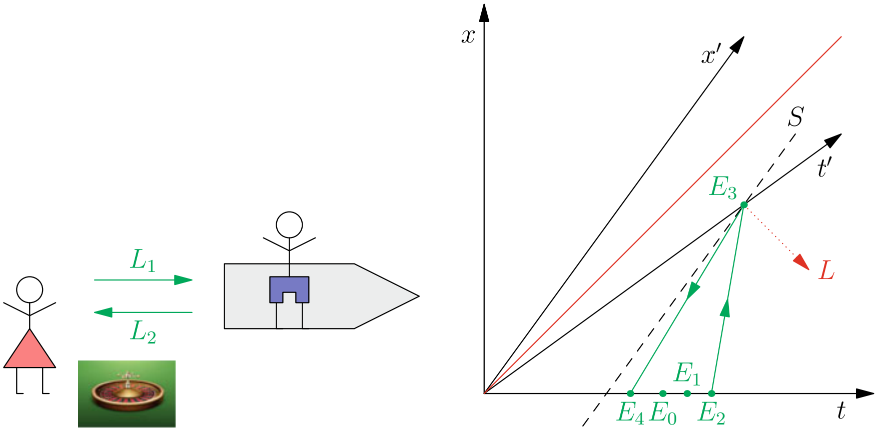


Fig. 7.11 The casino fraud

7.8 Causality and Faster-than-Light Velocity

7.8.1 The Casino Fraud

Alice and Bob have found a method that will allow them to always win in roulette. Alice is situated in the casino and Bob travels very fast relative to her. Both are in an inertial frame (see Fig. 7.11). Alice places bets. Then, the croupier spins the wheel and throws the ball into the wheel. Once the ball falls into the wheel, the croupier sweeps away the losing chips and makes the payouts.

Suppose that, at the event E_0, the croupier says “rien ne va plus” and no bets can be placed anymore. The ball is thrown into the roulette, and at event E_1, it comes to a rest in one of the pockets and the outcome is defined.

At the event E_2, just after the outcome is defined, Alice sends a message, which, for her, is faster than light, to Bob. The message arrives at Bob at event E_3. For Bob, the message travels backwards in time![16] The comparison of the line segment $\overline{E_2 E_3}$ with Bob’s x'-axis shows that $t'_3 < t'_2$.

Now, Bob sends the message back to Alice (for him, faster than light), where it arrives at event E_4.[17] For Bob, we have $t'_3 < t'_4$, but for Alice, Bob’s message travels backward in time, because $t_4 < t_3$. As a whole, Alice and Bob have achieved that the message emitted at E_2 arrives at Alice’s location at E_4. This is remarkable because E_4 chronologically comes before E_2. To make the fraud possible, Alice and Bob

[16] Whatever that may mean …

[17] The dotted red half-line L would be a message with light speed and the dashed black line S a “line of simultaneity” for Bob. Each half-line that starts at E_3 and passes between L and S in the negative x-direction is, for Bob, faster than light.

can adapt the method such that E_4 happens before E_0 (the “rien ne va plus”). Then, already before placing the bets, Alice knows where the ball will stop in the wheel.

Even more dramatic is the following thought. Alice sends a message to a long-time friend of her father, back into the past (in the same sense as in the casino fraud). The message contains the plea to prevents her parents from meeting. In this way, Alice could prevent her proper existence. But if she does not exist, she cannot send a message to her father’s friend. Obviously, sending a message into the past is nonsense.

Sending messages faster than light stands in contradiction to the **principle of causality**. This *principle on cause and effect* states that an effect is always the consequence of some cause and the former always happens after the latter: first the cause, then the effect.

In our example above, the event E_2 (Alice sends a message to Bob) is cause to the event E_3 (Bob receives the message from Alice and sends a message back to her), while E_3 is the effect to E_2. The event E_3, on the other hand is cause to the event E_4 (Alice receives the message from Bob), while E_4 is the effect to E_3. We have the *causal chain* $E_2 \to E_3 \to E_4$. As we have seen, for Alice, $t_4 < t_2 < t_3$, so the message from Bob to Alice violates the principle of causality for her. For Bob, $t'_3 < t'_4 < t'_2$, so for Bob, the message from Alice violates said principle for him.

The possibility to send messages faster than light contradicts the principle of causality. For that reason, the speed of light is the maximum velocity for messages (or signals). Instead of sending a signal, Alice and Bob could transmit the message with a messenger. For that reason, messengers (and all other observers) are also prohibited from traveling faster than light.

Speed of light as the upper limit on velocities[18]: The speed of light is the maximum velocity of a signal. Two inertial observers (or objects) relative to each other cannot move faster than light.

This “speed limit” does not imply that velocities larger than the speed of light are not possible in general. Suppose that Alice and Bob are standing on the Moon, distant from each other, and observe Claire, who is on the Earth. Claire has a laser and points to the surface of the Moon. She can move this point from Alice to Bob easily with a velocity larger than the speed of light. The important point here is that, in this case, the principle of causality cannot be violated, because, in this way, no causal connection between the event when the point passes Alice and that when it passes Bob can be constructed. Another example is when Alice sends a light pulse in the positive x-direction and another one in the negative x-direction. Then, the difference velocity of these pulses is $2c$. But relative to an inertial frame, an object can never move faster than light.

[18] The relative velocity between two inertial observers must be smaller than the speed of light. It can be arbitrarily close to the speed of light, but cannot reach it. See also Sect. 13.4.

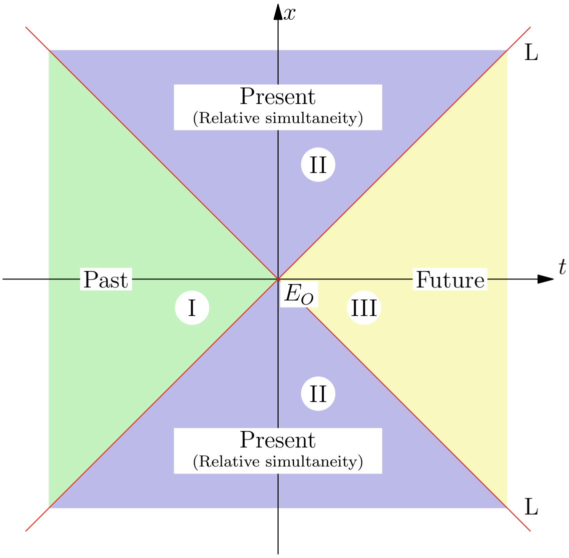


Fig. 7.12 The light cone L divides spacetime into different sectors with events that (I) lie in the past of the event E_O, (III) represent the future of E_O, and (II) are simultaneous to the event in the origin for certain inertial frames

7.8.2 *Past, Present and Future*

We come back to the two-dimensional spacetime diagram. Alice has the coordinate system (t, x) and Bob, as usual, travels relative to Alice with the constant velocity v and Bob with the coordinate system (t', x').

Consider a light signal that travels from the origin O in both directions in space. If we extend the rays (half-lines) of the light signal to negative times, we can divide spacetime into three sectors (see Fig. 7.12):

- **Sector I**, for which $t < 0$ and $|x| \leq -ct$ holds;
- **Sector II**, for which $|x| > c|t|$ or $x = t = 0$ holds;
- **Sector III**, for which $t > 0$ and $|x| \leq ct$ holds.

The light lines belong to the sectors I or III (with exception of the origin E_O, $x = t = 0$, which belongs to sector II). There is no overlap of the sectors.

What do Bob's possible coordinate systems look like (in dependence of his velocity relative to Alice)? We know that, for his velocity, $-c < v < c$ must hold.

It is clear that, independent of Bob's velocity, his t'-axis always comes from sector I, passes through the origin E_O, and continues into sector III. For *all observers*, the events in sector I lie in time before the origin event E_O. Therefore, sector I is called the **(absolute) past** of E_O. In the same way, for all observers, the events in sector III happen after the origin event E_O. Sector III is the **(absolute) future** of E_O.

The x'-axis that belongs to the t'-axis is constructed by reflecting the t'-axis at the line $x = ct$. All possible x'-axes therefore start in sector II with $x < 0$ and continue in sector III with $x > 0$. Bob's three possible coordinate systems are drawn in Fig. 7.13.

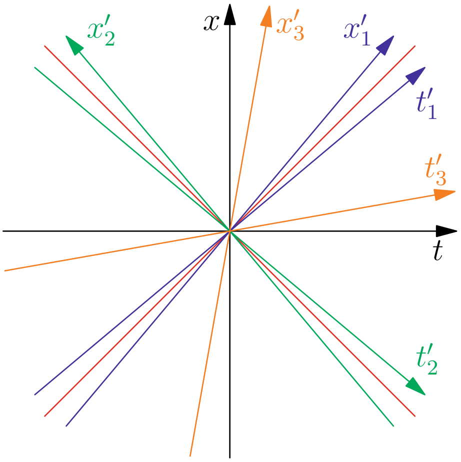


Fig. 7.13 Possible coordinate systems of moving observers

For each event E in sector II, one can find an observer, for whom the line passing through E_O and E is the x'-axis. For this observer, E_O and E are simultaneous. Sector II, for this reason, is the **(potential) present** of E_O.

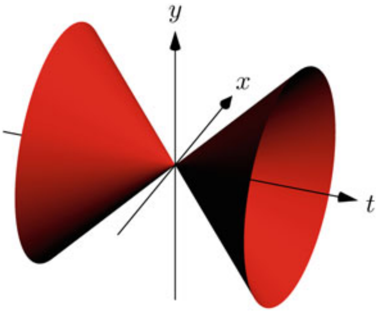
In the language of **causality**, each event in sector I can be the cause for an effect in E_O. The event E_O, on the other hand, can only be cause for effects in sector III. The event E_O can neither influence an event in sector II nor can it be influenced by those events, because the speed of light is the maximum velocity for signals (or information). The possibility that two events can be causally independent (i. e., they cannot influence each other) only exists in special relativity. In classical mechanics, there is no maximum signal velocity, and for two given events (if they are not absolutely simultaneous), one of them can always influence the other.

So far, we have taken only *one* space dimension into account and, accordingly, sent the light signal only in the positive and negative x-directions. If we consider two space dimensions, the light signal spreads circularly around the origin. If we add the time dimension, we end up with a spacetime diagram like that in Fig. 7.14. The world line of the light signal becomes the surface of a cone (see Fig. 7.14). This is where the term **light cone** comes from. It is also used in a two- or four-dimensional spacetime.

If we consider all three space dimensions, the light signal spreads spherically around the origin in all space directions.

The division of spacetime into past, present, and future also works in three- or four-dimensional space time. In three spacetime dimensions, past and future are the cones (interior and surface). The present is everything outside the cones (plus the origin). The case for four spacetime dimensions is difficult to illustrate.

Fig. 7.14 Light cone in three-dimensional spacetime



7.9 Digression: Rotating Reference Frames

7.9.1 Again: Synchronization of Clocks

In Sect. 7.3, we have given a procedure for synchronizing a clock B according to another clock A. Let the distance between the clocks be D. Then, when clock A shows time t_A, we send a light pulse to the other clock. Once this clock receives the light pulse, we set it to the time $t_A + D/c$.

Such a procedure only makes sense if it is free from contradictions. For this to be the case, it has to fulfill certain requirements. One of these requirements is called **transitivity**. Suppose we have three clocks and we want to synchronize two of them, B and C, according to clock A. To achieve this, we first synchronize clock B according to clock A, and then clock C according to clock A. Thus, we expect that clocks B and C are automatically synchronized. If this is not the case, the clocks, either with our procedure or even in general, cannot be synchronized. In inertial frames, clock synchronization with the given procedure is always feasible and no contradictions arise.[19]

However, in non-inertial frames, a clock synchronization in general is not possible, and the following example (see Fig. 7.15, left side) demonstrates this.

Our task is to synchronize a clock B according to a clock A. Both are located on the equator, clock A at location P_A and clock B at location P_B. The locations are antipodal regarding the center of the Earth. In order for us to be able to use our standard synchronization procedure, there is one fiberglass cable going from clock A westwards along the equator to clock B and another one going eastwards. Both fiberglass cables have the same length $l = \pi R_{\text{E}}$ (R_{E} is the Earth's radius). In this way, we have two possibilities to synchronize the clocks: westwards or eastwards from clock A to clock B.[20]

[19] We assume that the influence of gravitation can be neglected.

[20] As we already mentioned, for the standard synchronization procedure, we assume that the light pulse travels with the speed of light. This is not true in fiberglass, for which the speed of light in an inertial frame typically is about 2/3 of the speed of light in vacuum. For our discussion, this is not a problem, as we could use mirrors to keep the light pulse going around the Earth.

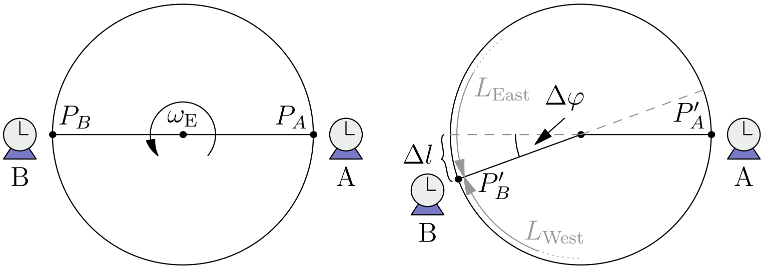


Fig. 7.15 Applying the synchronization procedure to two clocks A and B that are located on the equator

To describe the synchronization, we use an inertial frame in which the Earth's center of mass rests and that does not rotate with the Earth (e. g., the ecliptic geocentric coordinate system, egcs, see Sect. 4.4). In this inertial frame, the Earth rotates around its axis approximately once in a day (see Fig. 7.15, right side).

First, we use the fiber going eastwards (counterclockwise in the figure) from clock A to clock B. At time t_A, we emit a light pulse at P_A into the fiber. The light pulse travels along the equator to clock B. But due to the Earth's rotation, clock B moves away from the light pulse and, when the light pulse arrives, will not be at P_B anymore, but at another location P'_B. The light pulse has to travel the additional distance Δl. It arrives at clock B at $t_B = t_A + l/c + \Delta l/c$. Our synchronization procedure requires that we set clock B to $t_A + l/c$, because the light travels through a fiber of length l. As a consequence, clock B, in the inertial frame, is slow by the time $\Delta l/c$.

Now, we synchronize with the fiber going westwards (clockwise in the figure). Again, at t_A, we emit a light pulse into the fiber. This time, however, clock B travels toward the light pulse and, to reach clock B, has to travel the distance Δl, less than half of the Earth's circumference l. Our synchronization procedure now causes the synchronized clock B to be fast by the time $\Delta l/c$.

Depending on the path that we use to synchronize the clocks, we get different results. The synchronization procedure in rotating reference frames is not free from contradictions. In the case above, we could still synchronize the clocks in the inertial frame, which works fine. If there is an additional gravitational field, however, this, in general, is not possible anymore.

The time difference for clock B and the two synchronization paths is $\Delta t_B = 2\Delta l/c$. Now, Δl is the distance that a point on the equator moves while the light travels halfway around the equator. We have $\Delta l = v_{\text{ER}} \cdot l/c$, where $v_{\text{ER}} = 2\pi R_{\text{E}}/1\,\text{day} \approx 0.46\,\text{km/s}$ is the rotation velocity of the Earth at the equator. Therefore, $\Delta t_B = 2\Delta l/c = 4\pi^2 R_{\text{E}}^2/(c^2 \cdot 86{,}400\,\text{s}) \approx 200\,\text{ns}$. Whether we synchronize westwards or eastwards makes a time difference of about 200 ns. In this time, light travels about 60 m.

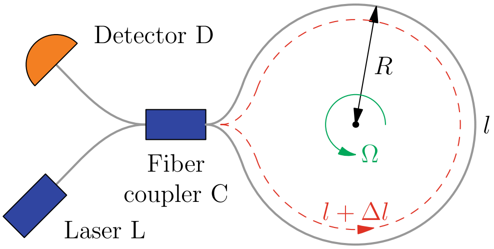


Fig. 7.16 Structure of a Sagnac interferometer

7.9.2 *The Sagnac Interferometer*

From the observations in Sect. 7.9.1, it also follows that the time needed by the light pulse to travel a closed path in a rotating reference frame, in general, depends on the direction of travel. This is the **Sagnac effect**. In the **Sagnac interferometer**, this effect is used to measure angular velocities of rotations.

The structure of a Sagnac interferometer in the easiest case is shown in Fig. 7.16. A laser L introduces a light beam into an optical waveguide, typically fiberglass. The light beam is split into two partial beams in a fiber coupler C, exactly as in a beam splitter (see Fig. 5.1). One partial beam travels clockwise and the other counterclockwise through the closed fiber loop with radius R and length $l = 2\pi R$. Each of the partial beams, after a certain time, arrives again at the fiber coupler only to become split into two further partial beams, leading to four partial beams. Two of them travel back to the laser; we are not interested in those (see Footnote 1 in Sect 5.2.1). The other two travel to the detector D. On the way to the detector, these two partial beams, which traveled through the fiberglass loop in different directions, interfere with each other. The intensity that is measured by the detector then depends on the phase difference $\Delta\varphi$ of the two partial beams.

Suppose that the whole setup, described from an inertial frame, rotates with an angular velocity Ω counterclockwise around an axis that is perpendicular to the fiber loop and passes through its center. Then, as seen from the inertial frame, the light waves travel different lengths, and this leads to a change in the phase difference $\Delta\varphi$. If we measure this phase difference, we can determine the angular velocity Ω. The Sagnac interferometer is a device for measuring the angular velocity.

How exactly does the phase difference depend on the angular velocity? Consider first the light wave that travels counterclockwise through the fiber loop. During the time $\Delta t = l/c = 2\pi R/c$ that a light signal needs to travel though the loop, the loop rotates by an angle of $\Omega\Delta t$. For this reason, the light wave, before arriving at the fiber coupler, has to travel an additional distance $\Delta l = R\Omega\Delta t = 2\pi R^2\Omega/c$. Using the area $A = \pi R^2$, this can be written in the form $\Delta l = 2A\Omega/c$. For a wavelength λ (in the fiber), this implies a phase difference of $\Delta\varphi_{\text{CW}} = 2\pi \cdot \Delta l/\lambda$, or

$$\Delta\varphi_{\text{CW}} = \frac{4\pi}{c}\frac{A}{\lambda}\Omega. \tag{7.4}$$

While the light wave traveling counterclockwise has to travel a distance of $l + \Delta l$, the light wave traveling clockwise only has to travel a distance of $l - \Delta l$, making for a phase shift of $\Delta\varphi_{\text{CCW}} = -\Delta\varphi_{\text{CW}}$. In total, we have a phase shift of $\Delta\varphi = \Delta\varphi_{\text{CW}} - \Delta\varphi_{\text{CCW}} = 8\pi A\Omega/(c\lambda)$.

At the detector, the two waves $a\sin(\omega t + \Delta\varphi_{\text{CW}})$ and $a\sin(\omega t + \Delta\varphi_{\text{CCW}})$ interfere with each other. This yields a total amplitude of $2a\cos(\Delta\varphi/2)\sin(\omega t)$. The detector sees an averaged intensity of $\bar{I} = 2a^2\cos(\Delta\varphi/2)$. Therefore, with an intensity measurement, one can determine the phase shift $\Delta\varphi$ (up to a multiple of 2π). From the phase shift via (7.4), the angular velocity Ω follows.

A Sagnac interferometer with a high sensitivity obviously needs a large area A and a small wavelength λ. One can multiply the area by using a fiberglass solenoid instead of a single loop. In the case of a solenoid with N windings, we have to replace the area A with $N \cdot A$ in the formula above and the sensitivity of the Sagnac interferometer is increased by a factor of N.

Exercise 22: You plan to travel to the north pole to measure the angular velocity Ω of the Earth's rotation using a Sagnac interferometer. To get a good resolution, the interference pattern should move by $\pi/10$, i. e., a tenth part of the wavelength.[21] Suppose the wavelength is 650 nm. What is the needed product $N \cdot A$ of the winding number and the area of the Sagnac interferometer?

Exercise 23: Approximately five minutes after one o'clock, the hour and minute hands of a clock point in exactly the same directions. Calculate the exact time when this happens. Discuss how this relates to the explanation of how the Sagnac interferometer works. What is the exact value of the length Δl for the Sagnac interferometer?

[21] To determine the phase shift, one can measure it first in an inertial frame and then in the rotating reference frame. For the problem at hand, it is not possible to stop the Earth's rotation. But we can "reverse" the Earth's rotation by just reversing the Sagnac interferometer.

Chapter 8
Length Contraction

8.1 Introduction

If the relativity of simultaneity was already surprising, length contraction is even more so.

In this chapter, we will come to the conclusion that the length of an object (for a given observer) depends on its velocity: fast moving objects are shorter than objects at rest. This is a direct consequence of the relativity of simultaneity. As application, we discuss cosmic particles that enter the atmosphere much deeper than expected and ladders that are longer than a garage, but fit into it anyhow.

8.2 Derivation

8.2.1 Length Measurement

In the last chapter, on relativity of simultaneity, it was of central importance to define what "simultaneous" exactly means and how distant clocks are synchronized. It is now equally important to define exactly how lengths are measured. In this way, we start by clarifying what the notion of the "length" of a moved object really means.

Thus: how is the length of a moving object measured? Take a rod that moves along a coordinate axis (or along a scale) and is also oriented along it. Obviously, to measure its length, we have to determine where the front end and the rear end of the rod are on the coordinate axis (or on the scale) **at the same time**. Note that, for objects at rest, the condition "at the same time" is not required.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_8

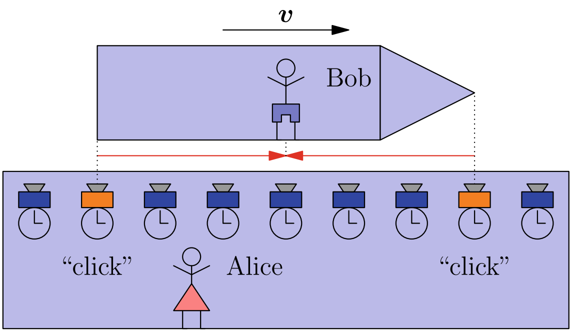


Fig. 8.1 Alice's cameras. The red arrows represent light pulses emitted by the cameras and for use by Bob

Rule for measuring the length of moving objects: To measure the length of a moving object, its front and rear ends are marked **at the same time** (for the measuring observer at rest) on a scale at rest.

Note that we do not move the scale with the object! The scale rests in the inertial frame of the observer.

8.2.2 Length Contraction in the Direction of Motion

Gedanken experiment. To demonstrate length contraction, we imagine Alice in the space station. At some point, Bob, who is sitting in the middle of his rocket, will fly by the space station with the velocity v relative to Alice, at which point Alice decides that she wants to measure the length of the rocket. To prepare for this, at each location of the space station, she places a device that consists of a camera, a clock and a light source (see Fig. 8.1). Then, she synchronizes the clocks of all these devices using the rule from Sect. 7.3. During measurement, the camera of each of these devices is triggered when the related clock shows a particular time. Thus, the cameras of all of these devices take a photograph *at the same time as far as Alice is concerned*. The devices are also able to process the photos immediately and to emit a light pulse when either the front end or the rear end of a rocket is displayed in a photograph. The camera that detects the rear end of the rocket sends a light pulse in the direction that the rocket is traveling and the camera that detects the front end of the rocket sends a light pulse contrary to the direction in which the rocket is traveling. Both Alice and Bob see the light pulses.

It is easy for Alice to determine the length of the rocket. She identifies the devices that detected the front end and the rear end of the rocket and measures their distance.

What about Bob? Bob is astonished that the light pulses emitted by Alice's devices do not arrive at his location at the same time. We also wonder, but know the reason: Bob moves toward one light pulse and away from the other. Because the front end and the rear end of the rocket are at the same distance from him, Bob concludes that the light pulses have not been emitted simultaneously (this is nothing but the relativity of simultaneity). He further observes that the camera at the rocket's front end took its photo earlier than the camera at the rocket's rear end. Bob understands that Alice, when measuring the length of the rocket, got a value that is too small, because the rear end of the rocket continued moving between the moments when the photos were taken. When Bob asks Alice what the length of his rocket is, he will indeed find that she gives a value that is too small. For Alice, the *moving rocket is shorter* than it is for Bob. Suppose Alice has the same rocket as Bob, but at rest. If she measures the length of this rocket, she will get exactly the same length as Bob provided her for the length of his rocket.

Alice now enters her rocket, which is the same model as Bob's rocket. Both fly by each other, Bob with the velocity v relative to Alice, as usual. We already know that, for Alice, Bob's rocket is shorter than it is for him. Because of the principle of relativity, the inertial frames of both are equivalent. *For Bob, Alice's rocket must also be shorter than it is for Alice.*

Even though Alice and Bob own the same rocket, Bob's rocket, *for Alice*, is shorter than her own rocket. Conversely, Alice's rocket, *for Bob*, is shorter than his own rocket. This may seem paradoxical (or even contradictory) to you, but is a necessary conclusion of the principles of special relativity. Space is relative! If the contraction were only to happen for one of the observers and a stretching the other, the principle of relativity would be violated. But the principle of relativity has been put to the acid test many times and has always passed it. The same holds for length contraction: it has been checked many times directly with experiments (see e.g., Sect. 8.3.1). There is no reason to doubt it. Length contraction is a *real effect* and not a kind of "optical illusion".

Due to the fact that the length of an object depends on its velocity, one should talk about the **proper length** of an object, which is the length of the object measured in its rest frame.

Graphical derivation. *How much* shorter is the moving rocket? Have a look at the spacetime diagram in Fig. 8.2. Alice and Bob now sit at the rear ends of their rockets. The spacetime diagram shows the trajectories of the front end and the rear end (t-axis) of Alice's rocket (in blue) and the trajectories of the front end and the rear end (t'-axis) of Bob's rocket (in green).

At time $t = 0$, Alice measures the length of her rocket and gets $l_0 = \overline{OE_A}$ (O is the origin of the coordinate system), which is the rocket's proper length. Furthermore, using our well-known rule, she measures the length of Bob's rocket and gets $l = \overline{OE_B}$. We know already that Bob's rocket, for Alice, is shorter than her own rocket, and therefore $l < l_0$. But we do not know how much shorter it is. For this reason, we have drawn the trajectory of the front end of Bob's rocket just such that it intersects the x-axis at a value l, which is little bit smaller than l_0. This "little bit smaller" has to be determined now.

Bob also measures the length of the rocket, namely, at time $t' = 0$. Remember that he has to measure the position of both ends of the rocket *at the same time for*

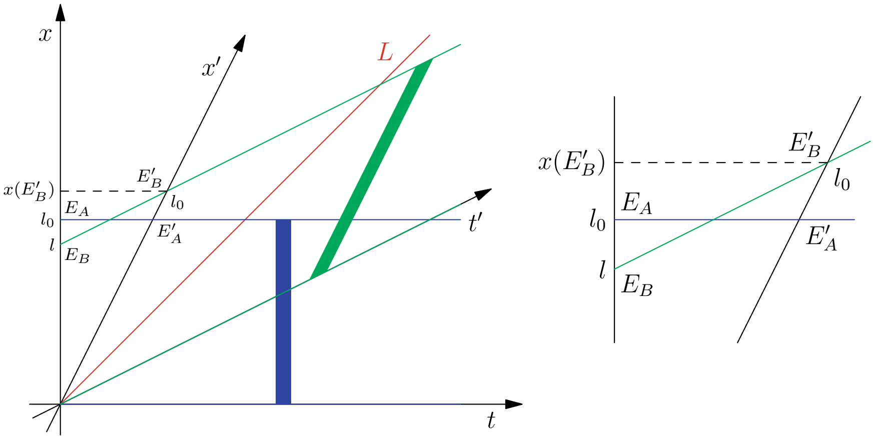


Fig. 8.2 On length contraction (see text) (Right: Enlarged section)

him. At time $t' = 0$, the rear ends of both rockets are at the origin of the coordinate system. The front end of Alice's rocket is at E'_A and that of his own rocket at E'_B. Thus, for him, the length of Alice's rocket is $\overline{OE'_A}$ and that of his own rocket $\overline{OE'_B}$. Because Bob has the same rocket as Alice, when measuring the length of his rocket, he gets the same result as Alice, when she measures the length of her rocket. Bob gets $l_0 = \overline{OE'_B}$ for the length of his rocket, which corresponds to the x'-coordinate of E'_B.

One notices already that lengths must not be transferred from Alice's to Bob's space axes (e. g., with a compass). The same holds for the time axes. Bob's axes have a *different scaling* than Alice's axes.

For Alice, Bob's rocket is shorter than her own. Because of the principle of relativity, for Bob, Alice's rocket must be shorter than his own by exactly the same factor. In other words: for Alice, Bob's rocket has the same length as Alice's rocket does for Bob. Therefore, we have $\overline{OE'_A} = l$. And consequently,[1]

$$\frac{l_0}{l} = \frac{\overline{OE_A}}{\overline{OE_B}} = \frac{\overline{OE'_B}}{\overline{OE'_A}}.$$

Because of the theorem on intersecting lines, we have

$$\frac{\overline{OE'_B}}{\overline{OE'_A}} = \frac{x(E'_B)}{l_0},$$

[1] As mentioned, Bob's coordinate axes have a different scaling than those of Alice. Lengths must not be compared directly in the spacetime diagram. What, however, can be compared are ratios of lengths or time intervals. This is because, in ratios, the scaling factors cancel each other out. Note that we assume that the coordinate transformations are linear.

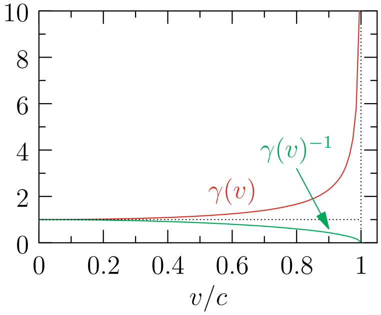


Fig. 8.3 The relativistic factor $\gamma(v)$ (red) and its inverse (green)

where $x(E'_B)$ denotes the x-coordinate of the event E'_B. In total, we infer that

$$\frac{l_0}{l} = \frac{x(E'_B)}{l_0} \quad \text{or} \quad x(E'_B) = l_0^2/l. \tag{8.1}$$

Now, we determine $x(E'_B)$ via the intersection of the trajectory of the front end of Bob's rocket with the x'-axis. Equating $x = l + vt$ and $x = (c^2/v)t$ and elementary transformations yields

$$x(E'_B) = \frac{1}{1 - v^2/c^2} l.$$

Equating this again with (8.1) yields

$$l = l_0 \cdot \sqrt{1 - v^2/c^2}. \tag{8.2}$$

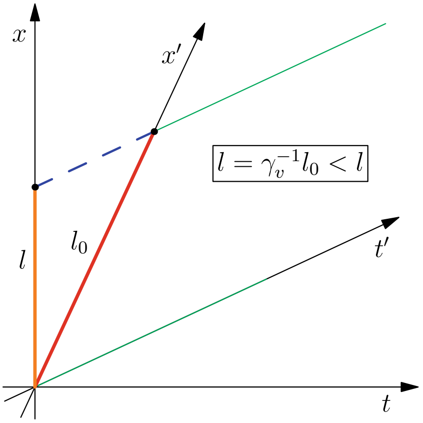
The γ-factor. We will encounter the factor with the square root quite often, therefore, we call it the **γ-factor** and write

$$\gamma_v = \frac{1}{\sqrt{1 - v^2/c^2}}. \tag{8.3}$$

The index v reminds us that this quantity depends on the relative velocity v. Sometimes, we more correctly write $\gamma(v)$, but the notation γ_v has the advantage of being short and the formulas keep being clear. Sometimes, we suppress the dependency on v completely and write simply γ instead of $\gamma(v)$.

What does the function $\gamma(v)$ look like? In Fig. 8.3, $\gamma(v)$ is drawn as a function of the velocity of the rest frame of the considered object. First, for $|v| < c$, we always have $0 \le v^2/c^2 < 1$, and therefore $0 < \gamma_v^{-1} \le 1$. For the inverse, $\gamma_v \ge 1$ holds. For $v/c \ll 1$, $\gamma(v)$ barely deviates from 1. This is the region in which classical mechanics is a very good approximation to special relativity, because, in classical mechanics, $l = l_0$, and therefore, necessarily, $\gamma(v) = 1$. If v comes close to the speed of light, $\gamma(v)$ raises strongly and eventually diverges at $v = c$.

Fig. 8.4 To length contraction



Conclusion. Equation (8.3) thus means that a moving object in the direction of motion is **shorter** when it is at rest and that the larger its velocity is, the shorter it is:

Length contraction: A moving object in the direction of motion is shorter by a factor of

$$\gamma_v^{-1} = \sqrt{1 - \frac{v^2}{c^2}} < 1.$$

than when it is at rest.

Or: An object that moves with velocity v and that has a proper length l_0 (in the direction of motion) has the length

$$l = l_0/\gamma_v. \tag{8.4}$$

For velocities v that are small in comparison to the speed of light, the length contraction is negligible. The larger the velocity v gets in comparison with the speed of light, the larger γ_v is and the smaller its length $l = l_0/\gamma_v$ becomes.

Discussion. We summarize the effect using Fig. 8.4. Bob travels with velocity v relative to Alice, and the standard configuration prevails. Bob has a rod[2] that rests in his inertial frame (red line in the figure) and has a proper length of l_0. The rod lies parallel to the x'-axis.

Now, Alice measures the length of the rod. According to the rule for measuring the length of moving objects, she measures the length of the orange line in the figure and gets the contracted length $l = l_0/\gamma_v$.

▌ **Exercise 24**: Calculate $\gamma(v)$ for $v/c = 0.5, 0.95, 0.995, 0.9995, 0.99995$.

[2] Here, we use a rod instead of a rocket. This is more down-to-Earth, in a double sense.

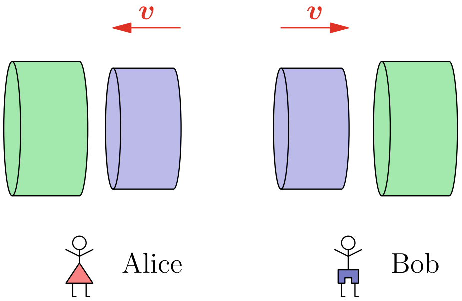


Fig. 8.5 Both Alice and Bob use a piece of pipe to investigate a possible length change transversal to the relative velocity

8.2.3 Digression: Length Change Transversal to the Direction of Motion?

Our previous investigations show that there is a contraction of moving objects in the direction of motion. Is there also a **length change transversal to the direction of motion**? This question can be answered with an easy Gedanken experiment. Imagine two short pieces of pipe with the same dimensions and an infinitely thin wall. One piece of pipe stays with Alice, the other one accompanies Bob on his travels. At some point, Bob reaches his traveling velocity and flies by Alice with the velocity v. Both have oriented their pieces of pipe in the direction of the relative velocity.

We know that Bob's piece of pipe, as seen from Alice's perspective, is shorter than her own piece of pipe by a factor of γ^{-1}. What about the diameter of the pieces of pipe? Suppose that it also changes, for instance, it contracts by a factor of $0 < \alpha < 1$ (see Fig. 8.5). We stay with Alice. She discovers that Bob's piece of pipe has a smaller diameter than her own piece of pipe. From her point of view, Bob's piece of pipe can fly perfectly through her own piece of pipe. What does Bob have to say? For him, Alice's piece of pipe has a smaller diameter than his own piece of pipe. From his point of view, Alice's piece of pipe can fly perfectly through his own piece of pipe. And herein lies the contradiction. The statement as to whether or not a piece of pipe can pass through another one and which one is smaller is a statement that cannot depend on the observer. Alice and Bob must come to the same conclusion in this case. For this reason, there cannot be a length change transversal to the direction of motion.

Remember the length contraction in the direction of motion. There, we said that, for each of the observers Alice and Bob, the other one's rocket is shorter than their own—and this despite the act that both own the same rocket. Here, we have a similar situation: each observer's piece of pipe is potentially smaller than the other's—and this despite the fact that both observers own the same piece of pipe. The relevant difference is that, in the latter case, we can compare the pieces of pipe directly. The statement that one of the pieces of pipe fits into the other one is not a *relative* observation but rather an *absolute* observation, i. e., equally valid for Alice and Bob.

Length contraction (continuation): A moving object does not change its length transversal to the direction of motion.

We will undertake a (failed) attempt to construct a situation in which two apparently comparable objects change their lengths in a different way in Sect. 8.3.2. And you will find a further reason why there is no length change (or contraction) transversal to the direction of motion in Exercise 32.

8.3 Examples

8.3.1 Muons

Cosmic radiation contains protons of very high energy. If these enter the upper layers of the atmosphere, they collide with nitrogen and oxygen atoms. As in particle accelerators, this causes showers of new particles, amongst them muons. Muons are elementary particles that are very similar to electrons, but heavier. However, muons are not stable. They decay with an half-life of about $t_{\mathrm{H}} = 1.5\,\mu\mathrm{s}$ into electrons and electron neutrinos.[3] One knows that these muons are produced at an altitude of about 10 km. Now, assume that the muons move at almost the speed of light relative to the surface of the Earth. Thus, after a flight distance of about $ct_{\mathrm{H}} = 660\,\mathrm{m}$, or at an altitude of more than 9 km, we should find only half as many muons. However, most of the muons still arrive at the surface of the Earth. How can this be explained?

The first experiment with cosmic muons was carried out by the American physicists Rossi and Hall in 1940 in the state of Colorado. A much more precise experiment of the same type was performed by the another set of American physicists, Frisch and Smith, in 1963 in the state of New Hampshire [FrischSmith63].[4] We will discuss the latter experiment. Frisch and Smith measured the number of muons that flow through a horizontal area per unit of time (see Fig. 8.6, left side). They performed the measurement twice: once at the top of Mount Washington and once in Cambridge (Boston) at sea level. On the mountain-top, they counted $N_1 = 563$ muons (within a certain time), and at sea level, $N_2 = 408$ muons (see Fig. 8.6, left side). The difference in altitude was $l = 1907\,\mathrm{m}$. Moreover, they measured the average velocity of the muons and got $0.995 \cdot c$.[5]

[3] This means that, if, at $t = 0$, you have a large number of myons that decay, then at $t = t_{\mathrm{H}}$, half of them will have decayed.

[4] At the time of publication, muons were called μ-mesons.

[5] Actually, their detector was built to detect only muons with a velocity between $0.9950 \cdot c$ and $0.9955 \cdot c$. For the detection at sea level, they also had to consider that the atmosphere slows down the muons.

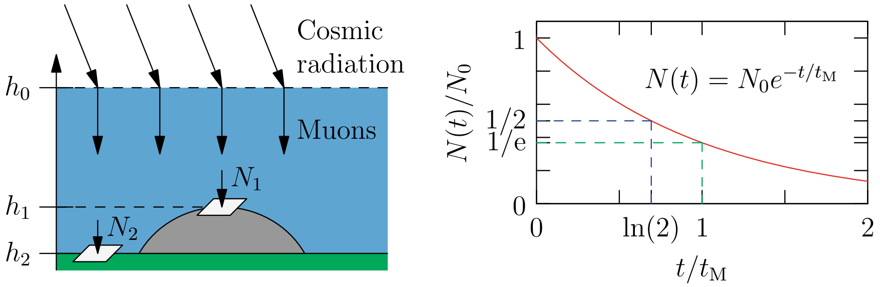


Fig. 8.6 Relativistic muons. Right: The principle of the experiment by Frisch and Smith. Left: The decay law with the half-life t_H

Let us take a look at the exponential decay (see Fig. 8.6, right side). Suppose that, at $t = 0$, in total, N_0 unstable particles *at rest* are given. In the course of time, particles decay such that, at time t,

$$N(t) = N_0 e^{-t/t_\mathrm{M}}$$

particles are left. The quantity t_M is the **average lifetime** of the particles. The average lifetime is related to the half-life via $t_\mathrm{M} = t_\mathrm{H}/\ln 2$. For the muons, we get $t_\mathrm{M} = 2.2\,\mu\mathrm{s}$.

We neglect the relativistic effects for the moment. Suppose the particles are created at $t = 0$ at the altitude of $h_0 = 10\,\mathrm{km}$ and travel perpendicular to and toward the surface of the Earth with the velocity v. Then, the instantaneous altitude is given by $h(t) = h_0 - vt$. At time t_1, N_1 muons have arrived at altitude $h_1 = h_0 - vt_1$, and at time t_2, N_2 muons have arrived at altitude $h_2 = h_0 - vt_2$. For the ratio N_2/N_1, one gets

$$\frac{N_2}{N_1} = e^{-(t_2-t_1)/t_\mathrm{M}} = e^{-(h_1-h_2)/(vt_\mathrm{M})} = e^{-l_0/(vt_\mathrm{M})} \quad \text{(non-relativistic)}, \tag{8.5}$$

where $l_0 := h_1 - h_0$ is the height of Mount Washington. In the experiment by Frisch and Smith, $N_2/N_1 = 408/563 = 0.725$, therefore, 72.5% of the muons that flew by Mount Washington arrive at sea level. According to (8.5), we expect

$$\exp\left(-\frac{l_0}{vt_\mathrm{M}}\right) = \exp\left(-\frac{1907\,\mathrm{m}}{0.995 \cdot c \cdot 2.2\,\mu\mathrm{s}}\right) = 0.055,$$

i.e., only 5.5% of the muons that pass by Mount Washington should arrive at sea level.

This is a huge gap. But it is also clear that, because of $v \approx c$, we have to take into account the effects of special relativity.

Let us put ourselves in the place of the muon that rushes toward the surface of the Earth. Then, Mount Washington rushes toward us with a velocity of $v = 0.995\,c$.

The distance of 1907 m from the top of Mount Washington to sea level is length contracted. Therefore, in formula (8.5), we have to replace l_0 with the contracted length l_0/γ. This yields

$$\frac{N_2}{N_1} = \exp\left(-\frac{l_0}{\gamma_v v t_{\mathrm{M}}}\right). \tag{8.6}$$

Because of $v = 0.995\,c$, we have $\gamma_v \approx 10$, and therefore $N_2/N_1 = 0.748$. Thus, the measured value for N_2/N_1 and the prediction using the decay formula while taking into account the length contraction agree to about 3%. In view of the difficulties of the experiment, this is a very good agreement. Taking into account the relativistic effect of length contraction, we can explain the result that Frisch and Smith got!

The observer on the surface of the Earth, however, cannot explain the experiment in this way, because, from his point of view, the distance from the location of the collision high up in the atmosphere to the top of Mount Washington (or sea level) is not contracted. Therefore, there must be another relativistic effect, which the observer on the surface of the Earth can invoke. This effect is the time dilation, and it is the topic of the next chapter. Time dilation obviously is needed, otherwise it would not be possible to explain the muon experiment from the point of view of both observers. Length contraction and time dilation are inseparable twins.

Exercise 25: What percentage of the muons that are produced at an altitude of 10 km actually reach sea level?

Exercise 26: Which value do we get from (8.6) with the slightly different velocity of $v = 0.993 \cdot c$?

8.3.2 *Ladder Paradox*

Next, we discuss a paradox of length contraction. It is a nice example of the intricacies related to the effects of special relativity and shows that a superficial consideration of special relativity may lead to apparent contradictions that vanish upon closer inspection.

Alice owns a garage (see Fig. 8.7, left side) which has a gate at each of the two opposite sides. The distance between the gates is the *length of the garage*. Alice also owns a ladder, which is a little bit longer than the garage. Now, Alice wants to place the ladder inside the garage. On first sight, this is not possible. But she remembers length contraction and accelerates her ladder to a velocity so high that it becomes shorter than the garage (see Fig. 8.7, center). She opens the gates and moves the ladder toward the garage. Once the ladder is in the garage, she rapidly closes the gates. Thus, at least for a moment, the ladder is in the garage, with the gates closed.

You already know what comes next. We take Bob's point of view, as he moves with the ladder. For him, the ladder is not length contracted, but the garage is, because it moves with the velocity $-v$ relative to Bob. For Bob, the ladder is definitely longer

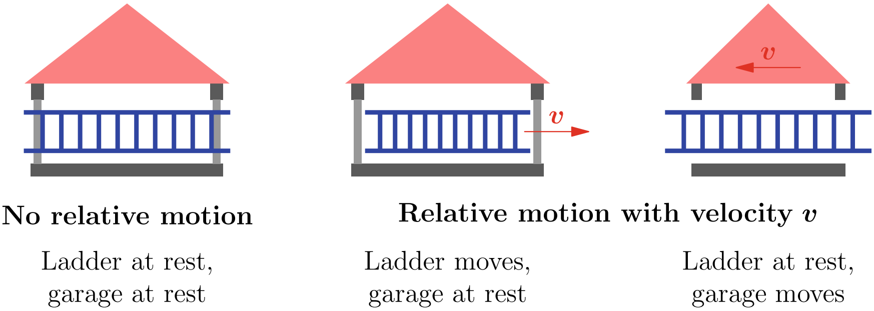


Fig. 8.7 The garage and the ladder. Left: Ladder and garage relative to each other at rest. Center and right: Ladder and garage move relative to each other. Center: Alice's point of view. Right: Bob's point of view

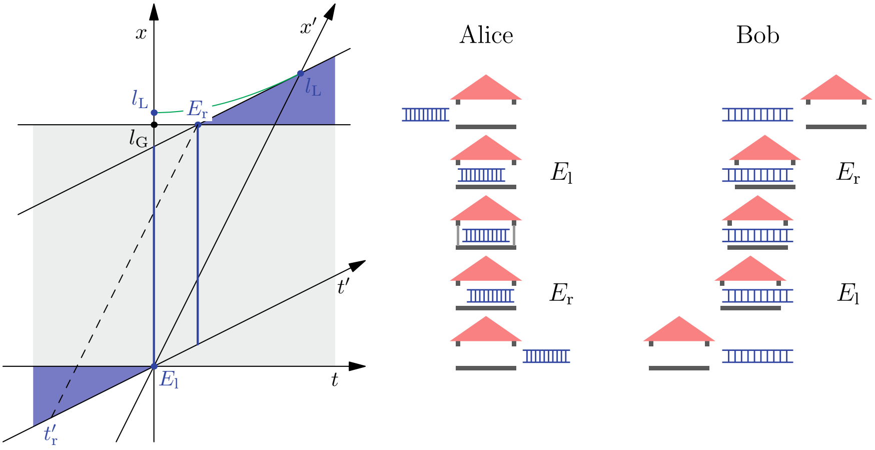


Fig. 8.8 Transit of the ladder through the garage. Left: Spacetime diagram. Center: Alice's point of view. Right: Bob's point of view

than the garage. From his point of view, it is completely impossible to place the ladder into the garage. Not even for a very small moment can both gates be closed simultaneously.

Who is right? Is there a moment at which the ladder is completely inside of the garage or not? In other words: does the moving ladder fit into the garage?

A look at the spacetime diagram in Fig. 8.8, left side, clarifies the matter. For Alice, the garage is at rest and has the proper length l_G, while the moving ladder has the length $\gamma_v^{-1} l_\mathrm{L}$, which, due to the length contraction, is smaller than the proper length l_L of the ladder. The proper length of the ladder, however, is longer than the proper length of the garage, $l_\mathrm{L} > l_\mathrm{G}$. Now, the ladder moves toward the garage with a velocity v large enough that $\gamma_v^{-1} l_\mathrm{L} < l_\mathrm{G}$ and enters it (see Fig. 8.8, center). In event E_l, the rear end of the ladder passes the *left* gate of the garage, and at event E_r, the front end of the ladder passes the *right* gate of the garage. For Alice, event E_l happens *before* event E_r, i. e., there is a small time interval in which, for Alice, the ladder is completely inside the garage. From Alice's point of view, the ladder fits into the garage.

For Bob (see Fig. 8.8, right side), the sequence of events is exactly the reverse. The front end of the ladder passes the right gate of the garage before the rear end of the ladder arrives at the left gate. For Bob, event E_l happens *after* event E_r, thus, for him, the ladder is never completely inside the garage.

The fact that the statement "the ladder fits into the garage" seems to be independent of the inertial frame (and therefore *invariant*), but is not, is responsible for the paradox. What does it mean that the ladder fits into the garage? It means that, when the ladder is in the garage, both gates can be closed **simultaneously**. In this definition, we encounter "simultaneity", a concept that is relative, as we have known since Chap. 7.

8.4 Digression: Hyperbolic Motion

8.4.1 Motion with Constant Acceleration

In our discussion of the ladder paradox in Sect. 8.3.2, there is a situation in which Alice can close the garage doors with the moving ladder completely inside the garage. What would happen if the ladder were then brought to rest? It must expand, but how?

This question brings us to accelerated motion, and we will discuss, in particular, the case of *constant acceleration*. To keep the discussion simple, we restrict ourselves, to one dimension, i.e., consider rectilinear motion only.

Let us start with the case of *constant acceleration in classical physics*. The acceleration of a moving body is invariant under Galilei transformations. If a particle moves with an acceleration $a = du/dt$ in one inertial frame, it has the same acceleration $a' = a$ in any other inertial frames. This follows directly from differentiation of the Galilean addition of velocities (which gives us the transformation of the velocity) with respect to time and the fact that $t = t'$.

In *special relativity*, this is no longer the case, and the acceleration of a particle is different in different inertial frames. The reasons for this are that the Lorentzian addition of velocities is no longer linear in the particle's velocity and that the time also transforms when changing inertial frames (see Exercise 28).

In classical physics, motion with constant acceleration is given by $x(t) = at^2/2$ and $v(t) = at$ and has the velocity increasing with time without limits. This is not possible in relativity, in which the speed of light is the limit velocity for particles and cannot be exceeded (or reached). For a fixed observer, there is no motion with strictly constant acceleration and that lasts forever. Therefore, we try to define a type of motion in special relativity that is similar in idea to the motion with constant acceleration in classical physics.

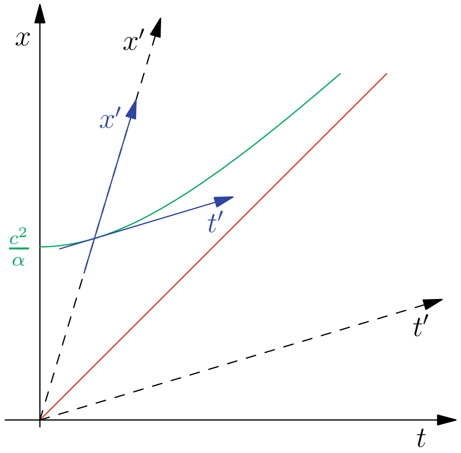


Fig. 8.9 Motion with constant proper acceleration

Suppose a particle starts in the regime of small velocities according to $x(t) = at^2/2$. Then, we can always go to the inertial frame in which the particle is momentarily at rest, the *instantaneous rest frame* of the particle, and apply a force to the particle that causes a constant acceleration. The acceleration of a particle in the instantaneous rest frame is called the **proper acceleration** α of it. If a box with a person inside is accelerated with a constant proper acceleration, then this person always feels the same acceleration and the person's accelerometer always shows the same acceleration value.

As we discussed, in classical physics, the acceleration of a particle is the same in all inertial frames and it does not make sense to define a *proper acceleration*. In special relativity, however, the proper acceleration α of a particle, in general, is different from the acceleration of the particle in other inertial frames.

The **motion with constant *proper* acceleration** in special relativity (for a fixed observer) is given by

$$x^2 - c^2t^2 = \frac{c^4}{\alpha^2} \quad \text{or} \quad x(t) = c\sqrt{t^2 + \frac{c^2}{\alpha^2}}, \tag{8.7}$$

which describes a *hyperbola* and is shown in Fig. 8.9, together with the coordinate system of the instantaneous rest frame of the particle.[6] This type of motion is called **hyperbolic motion**.

The transformation of (8.7) to a different inertial frame is rather easy. In Sect. 9.5, we will learn that the expression $c^2t^2 - x^2$ is an invariant regarding Lorentz transformations. Therefore, from (8.7), it follows directly that $x'^2 - c^2t'^2 = c^2/\alpha^2$. This tells us that the hyperbola looks exactly the same in all inertial frames, and this ensures that the proper acceleration, the acceleration in the instantaneous rest frame, is always the same for a particle moving according to (8.7).

[6] Note that this is for the initial condition $x(0) = c^2/\alpha$ and $v(0) = \dot{x}(0) = 0$.

It is easy to see that the velocity of the motion with constant acceleration never becomes equal to or larger than the speed of light, and for a fixed observer Alice, the acceleration decreases with time.

Exercise 27: Show that, for $\alpha t \ll c$, (8.7) becomes

$$x(t) = \frac{\alpha}{2}t^2 + \frac{c^2}{\alpha}.$$

Exercise 28:

- Differentiate the Galilean addition of velocities (3.5). With $a = du/dt$ and $a' = du'/dt'$, you get $a' = a$. In classical physics, the acceleration of a particle is the same in all inertial frames.
- Differentiate the Lorentzian addition of velocities (10.3). Show that, with $dt'/dt = \gamma_v \cdot (1 - uv/c^2)$, you get

$$\frac{du'}{dt'} = \frac{1}{\gamma_v^3 \cdot (1 - uv/c^2)^3} \frac{du}{dt}.$$

  This says that, if, in a coordinate system S with coordinates (t, x), a particle has the velocity u and the acceleration du/dt, then, in the coordinate system S' with coordinates (t', x'), the particle has the velocity $u' = u \ominus v$ (according to the Lorentzian addition of velocities) and the acceleration du'/dt' given by the formula above.
- Go to the instantaneous rest frame by using the Lorentz transformation with $v = u$. Show that the acceleration in this IS is given by

$$\alpha := \left.\frac{du'}{dt'}\right|_{\text{rest frame}} = \gamma_v^3 \frac{du}{dt}.$$

Exercise 29: An object in S moves according to the hyperbola $x^2 - c^2t^2 = c^4/\alpha^2$ and is instantaneously at rest at $t = 0$. Transfer the trajectory to the instantaneous rest frame I with coordinates (t', x'), of the object at t_0.

8.4.2 The Accelerated Rod

Suppose a rod (or our ladder) moves along the x-axis of Alice's coordinate system in the positive x-direction and is oriented parallel to this axis. The motion is such that each point of the rod is accelerated with the same acceleration a for Alice. Imagine that the rod is a measuring rod and each of its markings moves according to $x(t) = at^2/2 + x_0$, where x_0 is the initial position of the marking. For Alice, the

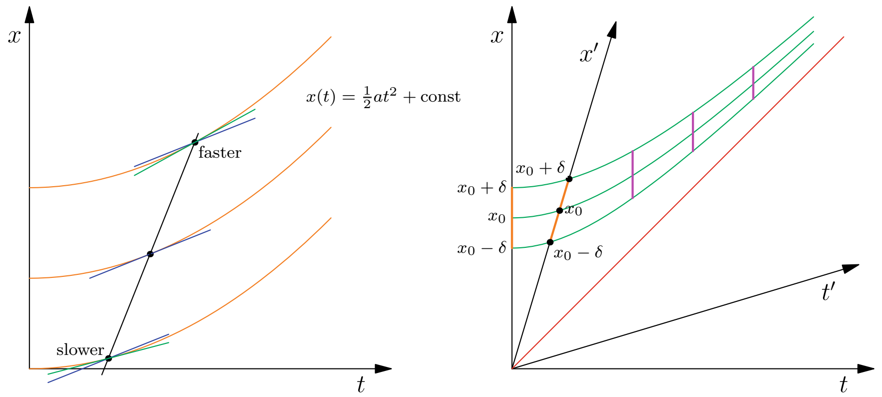


Fig. 8.10 The accelerated rod. Left: With constant acceleration a for Alice; Right: With constant proper acceleration α. Orange lines: Length is the same in all instantaneous rest frames. Magenta lines: Length becomes increasingly contracted for Alice

measuring rod then always has the same length. If x_L and x_R are the initial positions of the left and the right ends, respectively, of the measuring rod, then its length at time t is $x_R(t) - x_L(t) = x_R - x_L$ which is constant.

On the other hand, according to length contraction, the length of the moving rod should shrink! What will happen? Will the rod break?[7]

Yes, it breaks, and this can be seen in the following way. Take one marking on the moving rod, for instance, the middle marking, which initially is at position x_M, and go to the instantaneous rest frame I with coordinates (t', x') of this marking at some time t_0 (see Fig. 8.10 on the left). If we determine the velocity of the left and right ends of the rod in I, respectively, we see that both ends *move away from the middle marking*. The *reason for this is the relativity of simultaneity*. The left and right ends of the rod always have the same velocity at the same time for Alice. This is not the case in the instantaneous rest frame of the middle marking of the rod (neither is it for all other inertial frames that are not at rest relative to Alice). So, in the instantaneous rest frame I of the middle marking, the ends of the rod move away from the middle marking, the rod is stretched and will eventually break.

The interesting question now is: what kind of accelerated motion would keep the rod intact, without tension? Scrutinizing hyperbolic motion gives us the answer. In Fig. 8.10 on the right, we show the case where each of the rod's markings moves on a hyperbola of the form (8.7). Due to the fact that lengths are transferred from the x-axis of one inertial frame to the x'-axis of another one (see Sect. 9.5), the *length of the rod in the instantaneous rest frame is always the same* as the length of the rod

[7] The distinguished physicist John S. Bell, who is reponsible for one of the most important discoveries in quantum physics, made this Gedanken experiment famous. In his example, two rockets, at a certain mutual distance, were initially at rest in Alice's inertial frame and connected with a tight string. Then, the rockets were accelerated with an acceleration that was constant for Alice.

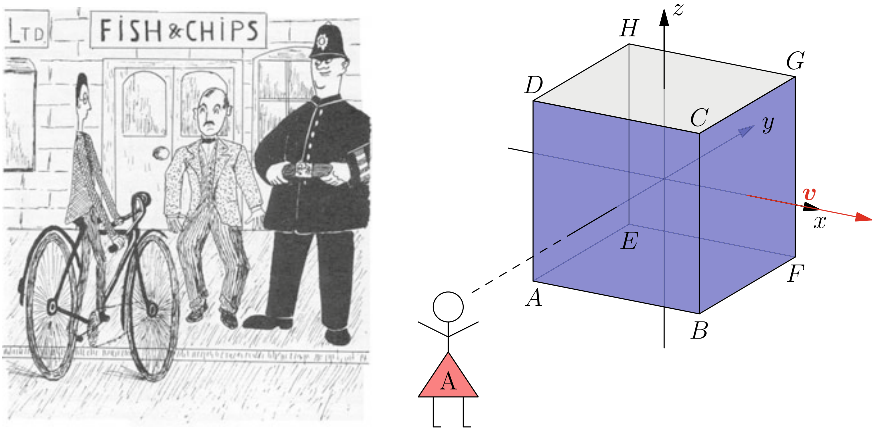


Fig. 8.11 Left: The length-contracted cyclist in Gamov's "Mr Tompkins in Wonderland". Right: Alice and her die

for Alice when it still was at rest for her. If each of the rod's markings moves on a hyperbola, the rod keeps its length and is not put under tension.

For Alice, however, the velocity and the acceleration of the right end of the rod are smaller than those of the left end of the rod. Again, this is to contract the rod's length in consistence with length contraction.

What does this mean for our initial question, bringing the moving ladder to a rest without breaking it? The answer is that each point of the ladder has to move on a hyperbola, which implies that it expands without tension to its rest length. While doing so, it will break open the garage's doors.

8.5 Visibility of Length Contraction[8]

Objects that fly by an inertial observer with a velocity v are contracted by a factor of γ_v^{-1} in the direction of motion while the dimensions transversal to the direction of motion stay the same. This we learned in Sects. 8.2.2 and 8.2.3, and this is exactly what the physicist George Gamov seems to depict in the illustrations in his popular science book "Mr Tompkins in Wonderland", published in 1940 (see Fig. 8.11, left side).

This depiction, however, is not completely correct, because there is a difference between what we get as a result when measuring the dimensions of an object and what

[8] This chapter was inspired by publications of the group led by H. Ruder from the University of Tübingen, Germany. A nice account is that by U. Kraus [Kraus+02]. In fact, the discovery that length contraction is invisible at not too high velocities because (under some circumstances) it is equal to a rotation was already demonstrated by J. Terrell in 1959 [Terrell59] and popularized by R. Penrose in the same year [Penrose59].

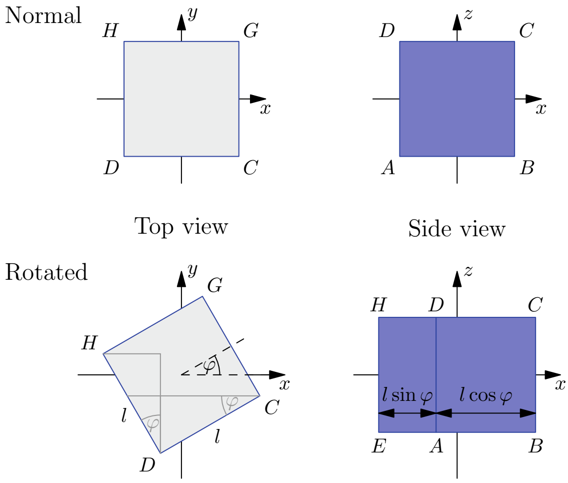


Fig. 8.12 Top: Normally oriented cube; Bottom: Rotated cube; Left: Top view; Right: Side view

we see when we look at it. The reason for this is the finite speed of light. Remember the comment about the bag of rice toppling over on the surface of Mars made at the beginning of Sect. 7.3? This is retardation! Indeed, a moving object does not appear contracted to the observer, but rather rotated and, in general, also distorted.

We discuss this effect by means of a cube with edge length l, which is oriented such that its center moves according to $x = vt$ in Alice's coordinate system and the edges are parallel to the coordinate axes (see Fig. 8.11, right side). The corners of the cube are denoted by $A, \dots, H$. Alice herself is located far away from the cube (such that effects of perspective play no role) and on the negative y-axis.

First, we **measure the cube** and use a measurement device very similar to that used in Sect. 8.2.1: on a plane $z = \text{const}$ just below the cube is a matrix of cameras with (for Alice) synchronized clocks. The cameras are all triggered at the same moment. Then, we seek the cameras that took photos of the lower corners A, B, E, F of the cube. We will see that the cameras that took photos of A and B (or E and F) are a distance l/γ_v apart, while those that took photos of A and E (or B and F) are a distance l apart. Indeed, the cube is contracted by a factor of γ_v^{-1} in the direction of motion, while it is unaltered in the transversal direction.

Then, we think about how the moving cube **would look for the observer Alice**, namely, at the time $t = 0$.

We consider first the **cube at rest**. In Fig. 8.12, top, on the left, the *top view* of the cube is shown and on the right, *Alice's view* is shown. If one rotates the cube by the angle φ around the x-axis, one gets the top view in Fig. 8.12, bottom left, and Alice's view in the same figure, bottom right. For Alice, the "apparent width" of the cube is $l \cdot (\sin\varphi + \cos\varphi) = \sqrt{2}l\cos(\varphi - \pi/4)$, which is larger than l if φ is not a multiple of $\pi/2$. Thus, the rotated cube, as seen by Alice, covers an area larger than l^2.

If Alice sees something like in Fig. 8.12, bottom right, and assumes that this is the projection of a rotated cube, she can calculate the rotation angle φ and the edge

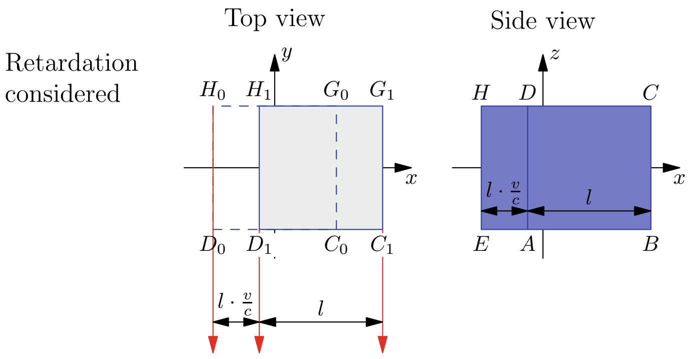


Fig. 8.13 Fast moving cube, taking into account retardation, *but not length contraction*

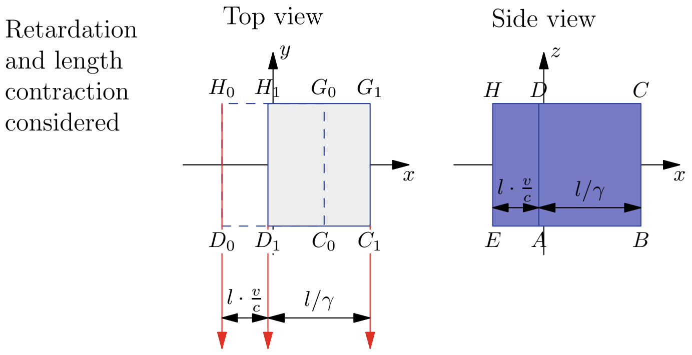


Fig. 8.14 Fast moving cube, taking into account retardation and length contraction

length $\tilde{l}$ of the cube in the following way: Alice lets $l_1 = \overline{AB}$ and $l_2 = \overline{EA}$, and thus the rotation angle is given by $\tan\varphi = l_2/l_1$ and the cube's edge length by $\sqrt{l_1^2 + l_2^2}$. In Fig. 8.12, bottom right, $l_1 = l\sin\varphi$ and $l_2 = l\cos\varphi$, and it is easy to see that this is consistent with our prescription above.

If **the cube moves with a high velocity**, we have to take into account the signal traveling times (or the retardation). In the first step, *we ignore the length contraction.* Figure 8.13 on the left side shows the top view of the cube. The position of the cube at $t = 0$ is indicated with dashed lines. The events corresponding to the cube corners are named C_0, D_0, G_0, and H_0. A time $\Delta t = l/c$ later, the cube has moved by a distance of $l \cdot (v/c)$, its new position is indicated with full lines and the events corresponding to the cube's corners are named C_1, D_1, G_1, and H_1. Now, the corner H is by a distance of l farther from Alice than the corner D. The light signal from H therefore needs the time l/c longer to arrive at Alice than that from D. Suppose a light signal from H and another one from D arrives at Alice's location at the same time. Then, the signal coming from H must have started at the time l/c earlier than the signal coming from D. During this time, the cube has moved by a distance of

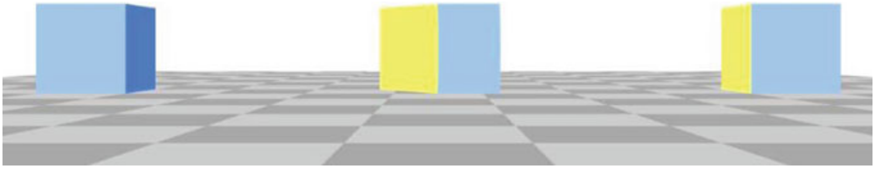
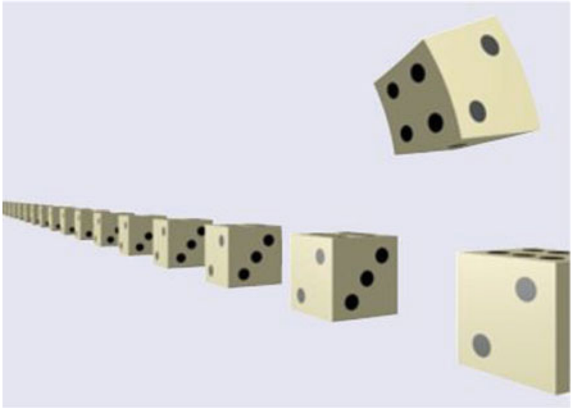


Fig. 8.15 Cube moving at 70% of the speed of light, as seen by a resting inertial observer

Fig. 8.16 Die that moves at 96% of the speed of light relative to the observer and a row of dice at rest



$l \cdot (v/c)$. Therefore, the signal from event H_0 arrives at Alice's location at the same time as the signal from event D_1. The signal from event C_1 also arrives at the same time. As a consequence, Alice does not see the corners H and D of the cube as being one exactly behind the other, but rather as shifted by a distance of $l \cdot (v/c)$ in the direction of motion of the cube. Thus, the cube, for Alice, looks as it does in Fig. 8.13, right side. The narrow green surface on the left side of the image is the left side of the cube, and it is not visible for Alice when the cube is at rest (or slowly moving).

The cube looks as if it had been rotated, but, upon a closer look, Alice finds that the edge length of the rotated cube would be $\sqrt{l^2 \cdot (v/c)^2 + l^2} = l\sqrt{1 + (v/c)^2} > l$. What Alice would see if only the projection were in place but there was no length contraction would not be consistent with a rotated cube of edge length l.

Now only the **length contraction** is missing. To include it, we just have to compress the width of the cube in Fig. 8.13, left side, by a factor of γ_v^{-1} to l/γ_v (as we did in Fig. 8.14 on the left side). The length contraction only acts in the direction of motion, while the lengths of the lateral sides in the y- and z-directions stay the same. Alice's view is shown in Fig. 8.14, right side.

For the apparent widths of the surfaces, we have $(l \cdot (v/c))^2 + (l/\gamma_v)^2 = l^2$, so, considering both the retardation and the length contraction, the image of the cube is consistent with that of a rotated cube with edge length l. The rotation angle is $\tan\varphi = l \cdot (v/c)/(l/\gamma_v) = \gamma_v \cdot (v/c)$. For small rotation angles φ, we have $\tan\varphi \approx \varphi$ and $\gamma_v \cdot (v/c) \approx v/c$, or, in total, $\varphi \approx v/c$. So, even for not very high velocities ("non-relativistic velocities”), the rotation angle is unequal to zero. For velocities close to the speed of light, we have $\gamma_v \cdot (v/c) \rightarrow \infty$, and therefore $\varphi \rightarrow \pi/2$. If the velocity of the cube approaches the speed of light, the rotation angle approaches 90°.

Figure 8.15 shows how, for an inertial observer not infinitely distant from the cube, it would look as if it were moving at a relative velocity of $v = 0.7c$ (perspective effects are taken into account). First, while the cube is approaching, one sees the front and the right side of the cube. Then, when the cube is closest, one sees the front side and, very clearly, the left side. Somewhat later, one sees less of the left side. The reason for this is that the cube’s velocity component perpendicular to the observer decreases very quickly.

In Fig. 8.16, the observer is approximately at the same altitude as a row of dice that rest relative to the observer. Above them, an equally oriented die moves with $0.95\,c$. One mainly sees the side with the value four, which is hidden on the other dice, and the cube is very distorted. The cause of these distortions is that the die is close to the observer and our approximation of the infinitely distant die does not hold anymore.

Chapter 9
Time Dilation

9.1 Introduction

It is not only the lengths of rapidly moving objects that change. For such objects, time also proceeds more slowly. We will find this out by sending two clocks out on a journey. Thereafter, we will learn why your twin, after returning from a long journey, is younger than you are. As applications, we will again discuss muons, but this time from a different point of view, and then satellite navigation.

9.2 Derivation

For moving objects, it is not only their length in the direction of the motion that is contracted. There is another effect: *from the observer's point of view*, time passes more slowly for moving objects than for objects at rest.

This effect is not completely unexpected. We have seen in recent chapters that, in special relativity, space and time share many more properties than in classical physics. At the end of Sect. 7.5.3, we discussed the fact that different events that happen at the same location for Alice don't do the same for Bob. And that, in special relativity, different events that happen at the same time for Alice don't do the same for Bob (relativity of simultaneity). For this reason, it seems likely that, in addition to the "space effect" of length contraction, there could be a concomitant "time effect", which concerns a change in the course of time.

We show now that this is indeed the case.

Time dilation. But first, a step back: what does it mean that time passes more slowly for moving objects than for objects at rest? Remember the synchronized clocks in the Volksgarten in Düsseldorf (in Sect. 7.3). Let Alice be at rest relative to this grid of clocks. Bob initially stands next to one of these clocks and carries a clock of the same type with him. His clock operates synchronously with the other clocks. Now, Bob moves with his clock rapidly towards a different clock within the clock

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_9

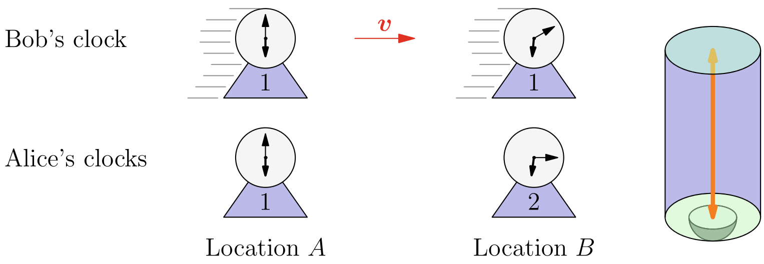


Fig. 9.1 Right: time dilation. Alice and her two synchronized clocks are at rest. Bob and his clock, which initially was synchronized with Alice's left clock, move relative to Alice. Left: Principle of a light clock. A sender/receiver (bottom) sends a light pulse upward. This light pulse is permanently reflected between the upper and lower sides of the clock. At the lower side, it is detected if it passes. The detection pulse is the ticking of the clock

grid. When he compares the time that his clock shows with the time of the clock next to him, he notes that his clock shows a slightly earlier time (the difference in time depends essentially on his speed, and in this case, it may be so small that the best atomic clocks would be needed to notice the difference). For Bob, the time passed more slowly on his journey from one clock to another in the Volksgarten. With this, we are talking about **time dilation**.[1]

Measuring the time. To measure the time, we use an (imaginary) **light clock** (see Fig. 9.1, right side).[2] A light clock consists of a hollow cylinder of length l, the two parallel planes of which are metalized on the inside. A light pulse, which travels along the cylinder axis, is then reflected permanently back and forth. Two reflections at one of the parallel planes happen with a period of $2l/c$. If one counts the number of reflections, one has a clock.

The light clock defines a unit of time as the time needed for a light pulse to travel a unit distance. This method works well for a clock, because, independently of the inertial frame and the location and direction, the speed of light is the same, and therefore the time needed for the light pulse to travel the unit distance l is always $t = l/c$. This is the statement of the absolute speed of light, and it was confirmed by the Michelson-Morley experiment. So, the location of the light clock and its orientation do not have an influence on its frequency.

Actually, however, the time is defined in a different way. Imagine you had a real light clock, a hollow cylinder of length l. The problem is that any material would change its length with temperature, and consequently the time measured by the light clock also would depend on the clock's temperature. As we will see in Sect. 9.3, in reality, periodic processes in atoms are used for time measurement. But conceptually, the light clock is a perfect fit for our purposes here, and Einstein's

[1] From the Latin *dilatare*, which means "enlarge" or "stretch".

[2] Note again that a light clock only makes sense in special relativity, because it has the same click frequency in all IS and for all directions. This is stated by the principle of the absolute speed of light and demonstrated by the Michelson-Morley experiment. In an aether theory, would not be true.

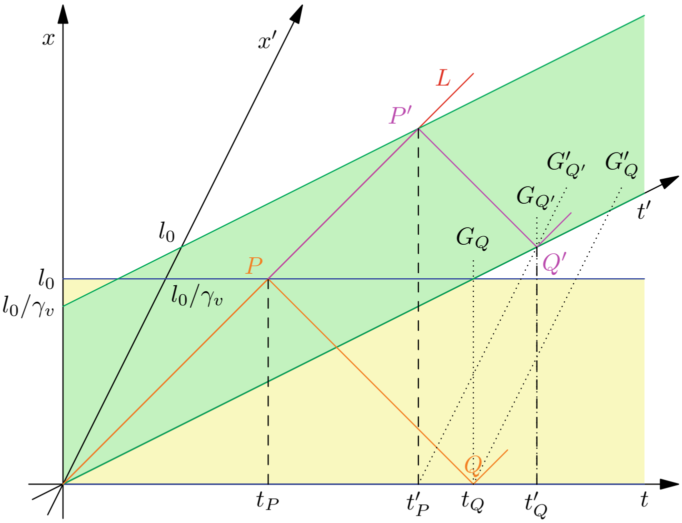


Fig. 9.2 To the derivation of time dilation (see text)

principle of relativity guarantees that an ideal light clock (with a fixed length, etc.) measures the same time as an atomic clock. Otherwise, comparing the pace of the two clock types, we could distinguish different inertial frames, in contradiction to Einstein's principle of relativity.

Graphical derivation. Alice and Bob again enter the scene. Bob moves with the velocity v relative to Alice. Both are inertial observers. And both carry a light clock, which they orient in the direction of their relative motion (i. e., the light pulse in the clocks travels parallel to the relative motion of Alice and Bob).[3] In the spacetime diagram (Fig. 9.2), the trajectories of Alice's light clock (the rear end corresponds to the t-axis and the front end to the blue line) and Bob's light clock (rear end = t'-axis, front end = green line) are shown. Both light clocks have the proper length l_0 (i. e., this is their length in the inertial frame in which they are at rest). We know already that, from Alice's point of view, Bob's light clock is length contracted by a factor of γ_v. Therefore, we put the front end of Bob's light clock on the x-axis at l_0/γ_v. For Bob, the situation with Alice's light clock is analogous.

Both light clocks start at the common origin O of the coordinate systems ($t = t' = 0$, $x = x' = 0$), where the rear ends of the clocks meet. In Alice's (Bob's) light clock, the light pulse is reflected at event P (P'), and at event Q (Q'), it has completed one period. Therefore, at event Q, Bob's light clock shows the same time as Alice's light clock at event Q'.

[3] Due to the fact that, in an inertial frame, the speed of light is independent of the direction, the orientation of the light clocks does not matter. We choose the direction such that the derivation at hand is as easy as possible.

The important observation is that, *for Alice*, Q' happens *after* Q (see the lines of simultaneity G_Q and $G_{Q'}$). *For Alice, Bob's light clock runs more slowly than her own*, because, if, on her own clock, one period has passed, Bob's clock will not yet have completed a period.

For Bob, however, the event Q happens *after* Q' (see the lines of simultaneity G'_Q and $G'_{Q'}$). *For Bob, Alice's light clock runs more slowly than his own*, because, if, on his own, one period has passed, Alice's clock will not yet have completed a period.

The effect is reciprocal, exactly as it must be according to the relativity principle, because both Alice and Bob reside in an inertial frame, and therefore are on an equal footing.

By how much does the time of a moving object run more slowly? This is now relatively easy to find out. We only need to determine the ratio of the time intervals from the origin to $t_{Q'}$ and from the origin to t_Q, respectively. We use Alice's coordinates. The latter is easy: $t_Q = 2l/c$.

Now, we turn to $t_{Q'}$. At event P', the green line and the light line intersect, therefore, $l/\gamma_v + vt_{P'} = ct_{P'}$. It follows that $t_{P'} = (l/c)\sqrt{(1+v/c)}/\sqrt{(1-v/c)}$ and $x_{P'} = ct_{P'}$. At event Q', the line from P' to Q' and the t'-axis intersect, therefore, $x_{P'} - c(t_{Q'} - t_{P'}) = vt_{Q'}$. It follows that $t_{Q'} = 2t_{P'}/(1+v/c)$ and, finally, $t_{Q'} = (2l/c)\gamma_v = t_Q\gamma_v$. Suppose that $t_Q = 1$ s. *For Alice*, Bob's clock shows one time unit only at $t_{Q'} > t_Q$. Thus, again, for Alice, Bob's runs more slowly than her own.

Therefore, we reach the conclusion (see Fig. 9.3):

Time dilation: For an inertial observer, the time of a moving object runs more slowly by a factor of

$$\gamma_v^{-1} = \sqrt{1 - \frac{v^2}{c^2}} < 1$$

as if it was at rest.

While, for the observer, the time Δt_0 has passed, for the object moving with velocity v, only the time

$$\Delta t'_0 = \gamma_v^{-1}\Delta t_0 < \Delta t_0 \tag{9.1}$$

has passed.

Discussion. In our graphical derivation, we have shown that, *for Alice, Bob's clock runs more slowly than her own clock*. In the first moment, one may not be terribly disturbed by this finding, because one automatically hesitates to carry this over to other processes that occur from Alice's point of view in Bob's inertial frame. But because of the relativity principle, if the clock runs more slowly, then all other processes in Bob's inertial frame also must run more slowly. This includes that, from Alice's point of view, Bob ages more slowly than she does. Time dilation is a *real effect* and not a kind of illusion.

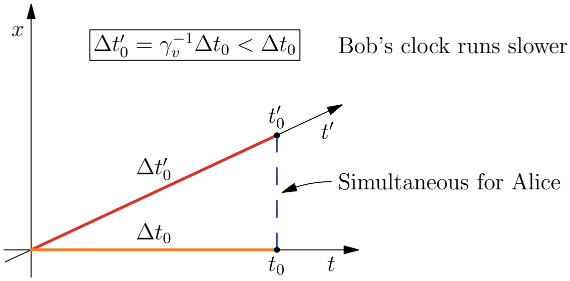


Fig. 9.3 To time dilation

Time dilation is an effect between two different inertial observers. From Alice's point of view, Bob ages more slowly. Bob, however, does not notice this. All of the processes, including his aging and the period of his clock, proceed in the same way: there is no reference relative to which Bob could conclude that something in his inertial frame proceeds more slowly or more quickly than anything else.

The effect is completely symmetric. This is also a consequence of the fact that all inertial frames are on the same footing. There are no preferred or special inertial frames. Therefore, if Bob looks at Alice's clocks, he will find that they run more slowly than his own clocks and he will also conclude that Alice ages more slowly than he does. This is not a contradiction, because Alice and Bob rest in different inertial frames, and this means that they possibly meet once in their lives, but never again. Therefore, they cannot directly compare their ages. We will change this situation in the discussion on the twin paradox in Sect. 9.9.

Notwithstanding the fact that the effect is symmetric, we have used different methods to derive the length contraction in Sect. 8.2.2 and the time dilation here. For the length contraction, we applied Einstein's principle of relativity and needed both perspectives, that of Alice and that of Bob, as the construction in Fig. 8.2 on the right side shows. In the derivation of time dilation, however, Alice's perspective was sufficient and we needed the light clock. But this was just for practical reasons. We could perform the derivation of the time dilation with the same method that we used for the length contraction.

Exercise 30: Adapt the method that was used to derive length contraction to time dilation.

Exercise 31: An airplane flies with a velocity of 1,000 km/h once around the Earth (whose circumference is 40,000 km). How much time has passed at the airport where the airplane started and landed? And how much time has passed for the airplane?

Two remarks: First, note that the airplane, strictly speaking, does not rest in an inertial frame, because it changes the direction of its velocity. This does not matter here. (See also the discussion in Sect. 9.10.) Second, in Sect. 9.11.2, you will see that gravity also has an influence on how fast the time passes. Here, you only have to take into account the dilation of time due to the relative velocity.

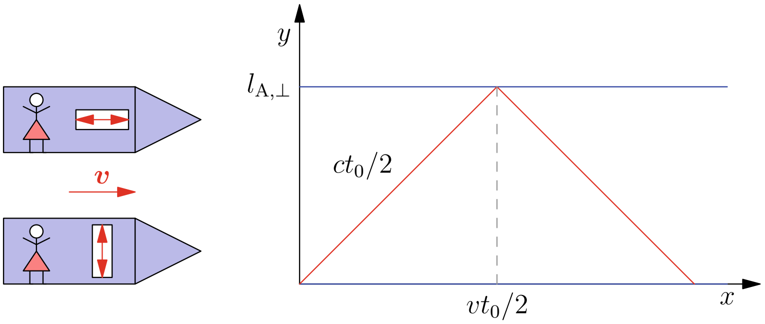


Fig. 9.4 Regarding Exercise 32. Left: Light clock oriented parallel to the direction of motion (in the longitudinal direction) and perpendicular to it (in the transversal direction). Right: Trajectory of the light pulse in the transversally oriented light clock. Caution: this is an x-y-diagram and not a spacetime diagram!

Exercise 32: Transversal and longitudinal light clock. In the derivation of time dilation, Bob's light clock was oriented parallel to his direction of motion (relative to Alice). Consider a light clock that is oriented perpendicular to the direction of motion (see Fig. 9.4, left side). Both light clocks must show the same time (argue why this is proved by the Michelson-Morley experiment). On the basis of this observation, argue that lengths are not contracted perpendicular to the direction of motion.

9.3 Digression: Time Measurement

Time measurement. The basis of time measurement is frequency measurement. You need something that oscillates at a fixed known frequency, the *oscillator frequency*, and then you count its cycles and convert them into seconds. A quartz crystal in a wristwatch, e. g., typically oscillates at a frequency of 32,768 Hz. The quartz clock counts the cycles (or periods) and, after 32,768 of these, one second has passed.

One of the first oscillators used for time measurement in modern history was the pendulum. You remember these **pendulum wall clocks**, like that in Fig. 9.5 on the left? The pendulum has a weight at the bottom whose distance to the axis of the pendulum can be changed and, in this way, the frequency of the pendulum can be adjusted to, e. g., one oscillation per second. For a wristwatch, the pendulum was not useful. Therefore, a spring-mass system was developed, the **balance-spring&wheel clock** (see Fig. 9.5 in the middle). These oscillators are still used in mechanical wristwatches.

All these oscillators have the same problem. The oscillation is **damped** and, at some point, it ceases to exist. Therefore, the oscillators have to be driven "from outside" to sustain the oscillation. In the case of a pendulum wall clock, this is a

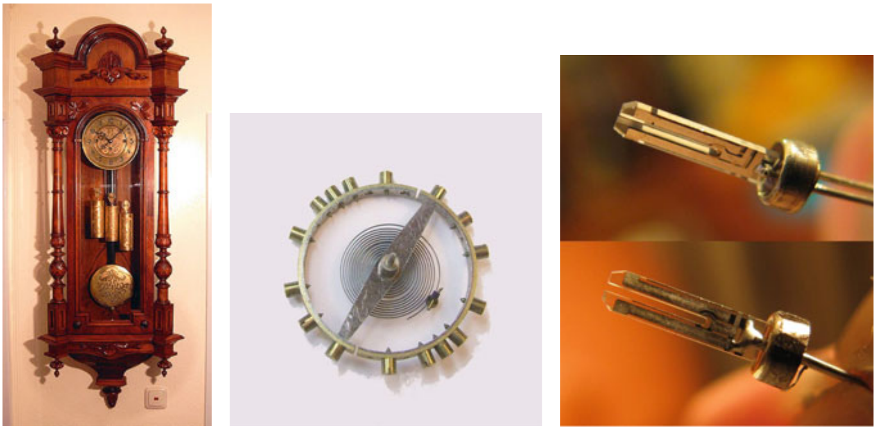


Fig. 9.5 Different types of clock oscillator. Left: Pendulum; Middle: Balance spring&wheel; Right: Quartz oscillator

weight fixed on a small chain (the pendulum clock in Fig. 9.5 has three of these). In the course of a day or so, the weight moves from its top to its bottom position and the pendulum clock has to be wound. This may be a minor nuisance, but it is related to a larger challenge. An oscillator that is damped necessarily also has **frequency noise**; its oscillation frequency is not fixed, but fluctuates *statistically* around the fixed oscillator frequency: sometimes it is a little bit larger, sometimes a little bit smaller. This means that a clock whose construction is based on a driven damped oscillator always has some error. And the larger the error is, the larger the damping is. The *Q factor* (quality factor) is a characteristic number of an oscillator that is inversely proportional to the damping on the one hand (the larger the damping, the smaller the Q factor) and to the frequency noise of the oscillator on the other. The higher the Q factor, the more *precise* the clock.[4]

There is a further source of errors. The oscillator frequency of such oscillators always depends on several external parameters. The most important one is usually the temperature. For instance, the balance wheel in a balance-spring clock, if not made with a very small thermal expansion coefficient, will change its size with the temperature, and this causes the clock to have a different oscillator frequency. Then the clock *systematically* runs too quickly or too slowly and will no longer be *accurate*.

Typical Q factors of good pendulum wall clocks or very good balance-spring clocks lie between 10^2 and 10^3. A big leap forward has been made with the introduction of the oscillating quartz crystals (see Fig. 9.5 on the right) used in **quartz clocks**. These have typical Q factors between 10^4 and 10^6.

[4] The two notions *precision* and *accuracy* are used to quantify measurement errors in general, as well as those of clocks. If all "seconds" of the clock have the same length, the clock is perfectly precise—even if the "seconds" are too long—but not accurate. And if the "seconds" all have a slightly different length but, on average, are exactly one second long, the clock is perfectly accurate but not very precise. Instead of precision, in the context of clocks, one also talks about their stability.

Much better Q factors can be achieved if we use atoms as natural oscillators.

Atomic clocks. Atomic clocks are our most precise time measurement devices. There are several reasons for this. First, some selected oscillations in an atom have a very low damping. Together with the fact that the oscillation frequencies usually are very high, this allows for a *very precise* frequency measurement. Furthermore, these oscillations in an atom are very robustly resistant to external influences, which leads to a *very high accuracy* of the clocks.

We will describe a typical atomic clock, the caesium beam atomic clock, in some detail in the next section. Basically, the caesium atom has two microscopic bar magnets, an electron and a nucleus, and these "rotate" about each other. The frequency of this rotation is very stable and accurate and has been used to define the second.

Definition of the second. The French *Bureau International des Poids et Mesures* is responsible for the International System of Units (SI). There, in 1967, the second was defined as:

> **Definition of the second**: A second is the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium-133 atom.

By the *transition between the two hyperfine levels of the ground state of the caesium-133 atom*, we mean the frequency of the mentioned rotation. In the clock, a microwave with a frequency of about $9{,}192{,}631{,}770\,\text{Hz}$ interacts with caesium atoms and, if the frequency exactly corresponds to the rotation of said microscopic bar magnets, one gets a resonance effect. Then, we are "only" left with counting the $9{,}192{,}631{,}770$ periods per second of this microwave and one knows that exactly one second has passed. This counting is performed with electronic counters.

Such an *atomic clock* is the *caesium beam atomic clock* CS2, located at the Physikalisch-Technische Bundesanstalt (PTB) in Braunschweig, Germany. Together with three further clocks,[5] the CS2 defines the official time for Germany. The clock has a precision of about 10^{-15}, i. e., in 100 millions of years, it accumulates, at most, 3 s of error. The clock is shown in Fig. 9.6. You can retrieve the current time via the Internet.[6]

Optical atomic clocks. Even better than an oscillator with its frequency in the microwave domain is one in the optical domain, and the reason for this is basically that the frequency of the transition is larger (in the case at hand, by a factor of several thousand). Thus, an even more precise time measurement is possible. Typical precisions are of more than 10^{-17}, which corresponds to a maximum error of 3 s in 10 billion years (for comparison: the Universe has an age of roughly 14 billion years). The fact that almost all atomic clocks these days still use oscillators with their

[5] Another caesium beam atomic clock CS1 and two *caesium fountain atom clocks* CSF1 and CSF2.

[6] See http://www.ptb.de/de/zeit/uhrzeit.html.

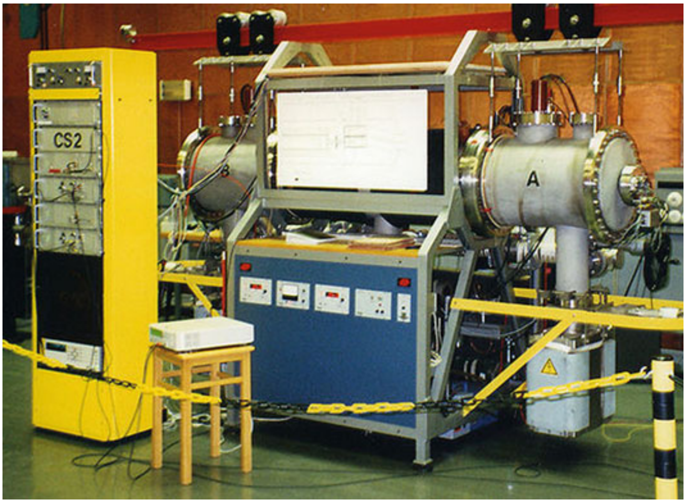


Fig. 9.6 The atomic clock CS2 of the Physikalisch-Technische Bundesanstalt (PTB) in Braunschweig, Germany

frequency in the microwave regime is for technical reasons. One of these reasons has been touched upon already: in the microwave domain, electronic counters can be used to count the clock's periods. In the optical domain, due to the high optical frequencies, this is not possible. One needs very complex *frequency combs* to divide the optical frequencies down into a regime that is accessible for electronic counters.

Using such an **optical atomic clock**, in 2010, the group led by David Wineland (one of the Nobel laureates of 2012) showed the time dilation for a relative velocity of only 10 m/s [Chou+10].[7] At this relative velocity, the difference between the clock rates of the resting and moving clocks is only about 5.6×10^{-16}. To demonstrate this, Wineland and colleagues compared this to the clock rates of a resting and a moving atomic clock and got a very good agreement with the predictions of special relativity.

Exercise 33: Show that the difference between the clock rates of a resting clock and a clock moving with a velocity of 10 m/s is 5.6×10^{-16}.

Exercise 34: What least amount of precision would an atomic clock need to have if its error must not exceed 1 s since the Big Bang?

[7] Time dilation was shown for the first time with an atomic clock in 1971 in a very famous experiment by Hafele and Keating, which we will discuss in Sect. 9.11.2. The two physicists flew together with four atomic clocks in an airliner around the Earth, and thereafter compared the displayed time with that of four further atomics clocks that had been left behind on the Earth's surface. See Sect. 9.11.2.

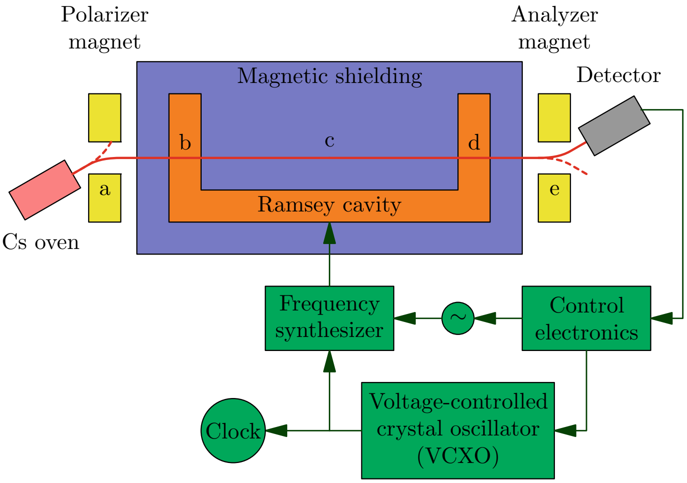


Fig. 9.7 Schematic view of a caesium beam atomic clock

9.4 Digression: Atomic Clocks

9.4.1 Overview

Figure 9.7 shows the composition of a **caesium** (Cs) **beam atomic clock**.[8] On the left side is a small *Cs oven* where caesium is vaporized. The caesium atoms then fly through a *polarizer magnet* (a) where the state of the Cs atoms is prepared. The atoms in the correct state then cross a first microwave cavity (b), fly freely for some distance (c), cross a second microwave cavity (d), and eventually the state of the atoms is measured in a combination of an *analyzer magnet* (e) and a *detector*. The green boxes represent an *electronic feedback loop* where, depending on the measurement outcome, the frequency of the microwave radiation in the microwave cavity (*Ramsey cavity*) is adjusted. If the measurement signal is maximal, the frequency corresponds to the Cs frequency standard and 9.192.631.770 cycles of it take exactly one second.

9.4.2 The Caesium Atom and Spin

For our purpose of understanding the Cs beam atomic clock, the Cs atom can be thought of as having one *electron*.[9] Furthermore, there is the *nucleus* of the Cs atom. In the stable isotope of Cs, which is what we consider here, this nucleus consists of

[8] For much more in-depth information on time standards and atomic clocks, see e. g., [Riehle04].

[9] Actually, the Cs atom has 55 electrons, but 54 of them are "paired" and have no influence on what we are discussing now.

133 nucleons (protons, neutrons) and is abbreviated by Cs-133. Both the electron and the nucleus have a **magnetic moment**: they act like two small bar magnets and dance around each other. This can also be described in an alternative way: the magnetic moment $\boldsymbol{\mu}_{\mathrm{e}}$ of the electron creates a magnetic field $\boldsymbol{B}_{\mathrm{e}}$ and the magnetic nucleus precesses in this magnetic field. Or: the magnetic moment $\boldsymbol{\mu}_{\mathrm{n}}$ of the nucleus creates a magnetic field $\boldsymbol{B}_{\mathrm{n}}$ and the magnetic electron precesses in this magnetic field. All of these are valid means to describe what happens.

Now, in the description of the physics of this system, we enter the strange world of **quantum physics** with *superposition*, *entanglement*, the weird effects of *measurement* and the like, but fortunately, for the *pas de deux* of the electron and the nucleus, there is a description that, on the one hand is exact (actually, it *is* the description according to quantum theory) and, on the other hand, is "minimally invasive" for our concepts of the world as digested from our day-to-day experience with the classical world. This description is that with the *Bloch vector*, which we present now.

Both the electron and the nucleus are *electrically charged*, and the reason for their magnetic moment is that they rotate around themselves and have an *angular momentum* like a spinning top. This angular momentum is called the **spin**.

The magnitude of the angular momentum (or spin) of both the electron and the nucleus is constant: they always rotate with the same angular velocity, but the rotation axis can change. Now, in quantum physics, we find the smallest non-vanishing spin, which is given by *half* of the fundamental physical constant $\hbar$, the **Planck constant**, which is very tiny.[10] The electron has this minimum angular momentum—it cannot rotate any more slowly—and particles with this minimum spin are called *spin-1/2 particles*. The nucleus of Cs-133 has a larger spin with value $7/2$ (times $\hbar$).

We will, however, treat our nucleus as a spin-1/2 particle. Then, the description becomes much easier without changing the physics, at least for the case of explaining how the Cs beam atomic clock works.

The spin state of a spin-1/2 particle (electron, proton) in quantum theory can be described exactly by a unit vector $\boldsymbol{e}_{\mathrm{S}}$, the **Bloch vector**, which can be considered as the particle's rotation axis. The unit vector also can be described by a point on the surface of the unit sphere, the **Bloch sphere**. Then, $\boldsymbol{e}_{\mathrm{S}}$ points from the center of the sphere to this point. So, the spin state of an electron or of a proton can be described by a point on the Bloch sphere. In this spin state, it rotates around the axis, going through the center of the Bloch sphere and this point.

[10] The German physicist Max Planck showed, in 1900, that to explain the spectrum of so-called black-body light, one needs to assume that energy is quantized. This was the discovery that triggered the development of quantum theory. The energy of a photon with frequency ν is given by $E = h\nu$, where $h := 6.62607015 \times 10^{-34}$ J/Hz is the Planck constant. For practical reasons, $\hbar := h/(2\pi)$ was introduced, and is also called the Planck constant. Thus, we can write $\hbar\omega$ instead of $h\nu$. In addition, spins are always multiples of $\hbar/2$.

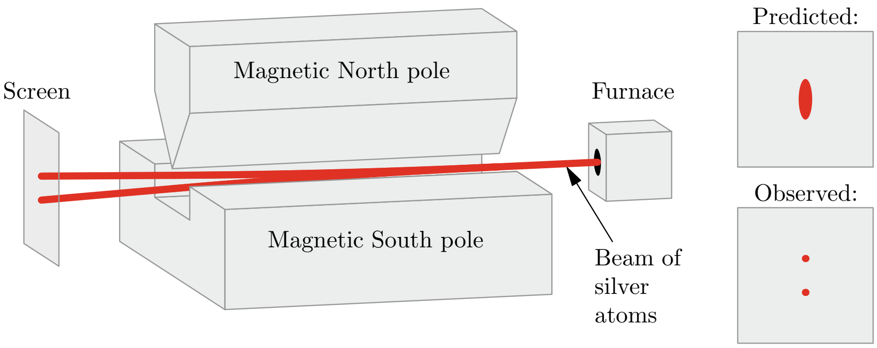


Fig. 9.8 Stern-Gerlach measurement device and experiment

9.4.3 Measuring the Spin

Enter the concept of **measurement** in quantum theory. As we stated, the electron (or the proton) can rotate around an arbitrary axis, given by the unit vector $\boldsymbol{e}_S$. The weird thing now is that we cannot *determine* this axis, but we still can *know* it! The reason for this is not that we do not have a suitable measurement device for determining the rotation axis, but rather that there is a general limit that Nature imposes.[11] What does this mean?

The prototypical spin measurement device is the **Stern-Gerlach (SG) device** (see Fig. 9.8). Central to the device is a strong inhomogeneous magnetic field produced by a magnetic pole pair. The objects whose magnetic moment is being detected (silver atoms in the original experiment by Stern and Gerlach, produced in a furnace) are sent through this magnetic field, where they are deflected depending on their magnetic moment and finally impinge on a detector screen (see Fig. 9.8, left side). The position where they are detected tells us about their magnetic moment.[12]

Little bar magnets. Suppose we have a very small bar magnet with magnetic moment $\boldsymbol{\mu}$, so small that we can barely see it, but too small to see in which direction its magnetic moment points. Then, we shoot this small bar magnet through the region with a strong inhomogeneous magnetic field $\boldsymbol{B}$ of the SG device (see Fig. 9.8, middle).

Let us orient the coordinate system such that the z-axis points in the direction in which this field changes maximally. Then, the force that the small bar magnet experiences is proportional to $\mu_z\, dB/dz$, the component of its magnetic moment in the direction of the (inhomogeneous) magnetic field, and the change in the magnetic field. This is equal to $\mu\, dB/dz \cos\varphi$, where φ is the angle between $\boldsymbol{\mu}$ and the z-axis and can take any value between $-\mu\, dB/dz$ and $+\mu\, dB/dz$. Accordingly, there is a line of possible locations where the bar magnet impinges on the detector

[11] You may have heard about *Heisenberg's undeterminacy relation*. This is what we are talking about.

[12] To be precise: about the component of $\boldsymbol{\mu}$ in the direction of the magnetic field.

screen. If we send many small bar magnets with random directions of the magnetic moment through the Stern-Gerlach device, we find this line on the screen (see Fig. 9.8, "Classical prediction").

Electrons. Let us turn from little bar magnets to **electrons**. If we send electrons through the Stern-Gerlach device,[13] something unexpected happens (see Fig. 9.8). Instead of a line, we see only two points on the screen (see Fig. 9.8, "Observation for silver atoms")![14] The electron is only found either deflected upward or downward, there are no intermediate deflection directions. This effect cannot be explained classically; it is a pure *quantum effect*. If the electron has been deflected upward, we say it has **spin up** with respect to the SG device's *measurement direction* (i. e., the direction of the magnetic field inhomogeneity) and its rotation direction is parallel to the measurement direction, $\boldsymbol{e}_S = \boldsymbol{e}_z$, and if it has been deflected downward, we say it has **spin down** and its rotation direction is antiparallel to the measurement direction, $\boldsymbol{e}_S = -\boldsymbol{e}_z$.

Nevertheless, we know (from interference experiments like that presented here) that the electron spin can point in any direction.

If we take the electrons that leave the SG device at its *spin-up exit* and pass them through a further SG device with the same orientation (or measurement direction), we always also find them leaving the second SG device at the spin-up exit and never at the spin-down exit. This means that all electrons leaving the SG device at the spin-up exit, independently of their earlier spin state $\boldsymbol{e}_\mathrm{S}$, have spin up (analogously for spin down), and consequently the act of measuring in general *changes the spin state* of the electrons.

In this way, the SG device can also be used to **prepare the spin state**[15] of particles, and this is performed by the *polarizer magnet* in the Cs beam atomic clock (Fig. 9.7), which is nothing but an SG device. If the spin state $\boldsymbol{e}_\mathrm{S}$ of the Cs atoms leaving the oven is random, then half of the atoms will leave the polarizer magnet at the upper exit with spin up (with respect to to the measurement direction) and will be discarded. The other half will leave the polarizer magnet at the lower exit with spin down and are used in the Cs beam atomic clock.

Measurement distribution and quantum randomness. Suppose we prepare electrons to have their spin up with respect to the y-axis by sending them through a SG device with the magnetic field inhomogeneity oriented in the y-direction and then selecting them. Next, we take these electrons and send them through an SG device

[13] This experiment would be challenging, because the charged electrons react much more strongly to stray electric fields than to the inhomogeneous magnetic field in the Stern-Gerlach device. Instead of electrons, Stern and Gerlach used (electrically neutral) silver atoms. These atoms have 47 electrons and, similar to the case of caesium, all of those but one are "paired". Therefore, the silver atom (without its nucleus) is also a spin-1/2 particle.

[14] In special cases, one of the two points may be absent. We will see later.

[15] This should resolve the puzzle introduced by the phrase, "The weird thing now is that we cannot *determine* this axis, but we still can *know* it!" If we don't know the electron's spin, we cannot find it out, because the measurement will change it in general. But after carrying out a measurement, we know the electron's spin state.

whose magnetic field inhomogeneity is oriented in the z-direction. We will find half of the electrons leaving the SG device at the upper exit and the other half at the lower exit. Whether a given electron leaves the SG device at one or the other exit is **completely random**,[16] but the probabilities of getting one or the other result (in this case, fifty-fifty) are fixed by the direction of the magnetic moment of the measured particles.

You could come up with the idea that, before the measurement, the state is actually always either spin up (with respect to the measurement direction) or spin down, but we just don't know. But this view is untenable, because, as discussed, we can prepare the spin to e. g., point in the y-direction and then measure it in the z-direction and will find 50% of the cases with spin up and the other 50% with spin down. But if we were to measure in the y-direction, we would always find it with spin up.

9.4.4 The Compound Particle

Compound spin. So far, we have learned how quantum theory describes a spin-1/2 particle like the electron or the proton. Actually, our Cs atom (with the fictitious spin-1/2 nucleus) contains two spin-1/2 particles, the electron and the nucleus, and both are relevant to the operation of the Cs beam atomic clock. Everything that we do with magnetic fields in the Cs beam atomic clock will act equally on the two spins. Fortunately, *for our purposes*, we can replace the two spin-1/2 particles with one artificial *compound particle* that can be described as a spin-1/2 particle with the Bloch vector $\boldsymbol{e}_{\text{en}}$ ("en" for "compound particle made of an *e*lectron and a *n*ucleus"). In this picture, the spin of the electron and that of the nucleus are always antiparallel to each other. If we want, we can interpret $\boldsymbol{e}_{\text{en}}$ as the spin of the electron and $-\boldsymbol{e}_{\text{en}}$ as that of the nucleus. This interpretation, however, is actually not correct, because the spins of the electron and the nucleus are usually *entangled* which means that only the whole composite particle has a spin while the constituents do not anymore.

Energy and free precession. Due to the magnetic interaction of the electron and the nucleus, the composite particle has a certain energy that depends on the state of its composite spin, i. e., in the direction of $\boldsymbol{e}_{\text{en}}$.[17] There is a certain direction $\boldsymbol{e}_{\text{en},0}$ with maximum energy E_1, thus $-\boldsymbol{e}_{\text{en},0}$ has the minimum energy E_0. If the composite spin takes one of these values and we measure the energy, we will always get the indicated value. In all other spin states, similar to the spin measurement, we will either get the maximum or the minimum energy, with certain probabilities.

[16] We talk about *quantum randomness*. Even Nature itself, which knows everything about the world before the measurement takes place and everything about the intricacies of the measurement device, is not able to predict the measurement outcome.

[17] Here, the picture of having one spin pointing in the direction $\boldsymbol{e}_{\text{en}}$ and the other in the opposite direction fails again, because the energy then would always be the same, independent of the direction. Due to the state superposition (or entanglement), this is not the case here. Remember that our description is exact; only the idea that we make out of it to "understand" it fails.

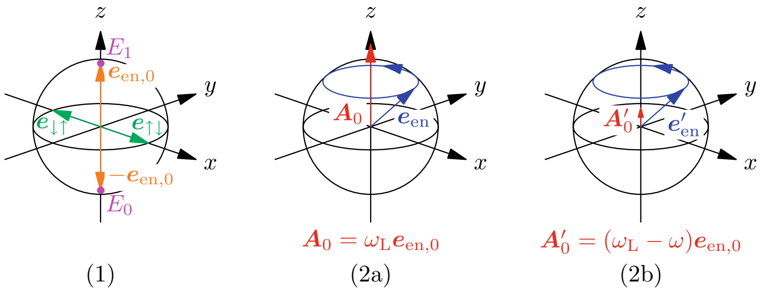


Fig. 9.9 (1) The Bloch sphere for compound spin; (2a) Free precession of the Bloch vector $\boldsymbol{e}_{\text{en}}$ around $\boldsymbol{A}_0$ in the laboratory frame and (2b) of the Bloch vector $\boldsymbol{e}'_{\text{en}}$ around $\boldsymbol{A}'_0$ in the rotating frame

The difference $E_{\text{HFS}} = E_1 - E_0$ between the maximum and the minimum energy is called the **hyperfine-splitting**[18] (HFS) **energy** in the ground state of Cs and, according to the definition of the second, is given exactly by the Planck constant h times the 9.192.631.770 Hz.

We introduce a coordinate system such that $\boldsymbol{e}_{\text{en},0}$ points in the z-direction (see Fig. 9.9(1)). As for a spinning top that precesses, the spin $\boldsymbol{e}_{\text{en}}$ of the composite particle is not static, but it rotates (precesses) around the direction given by $\boldsymbol{e}_{\text{en},0}$ (see Fig. 9.9(2a)). This is called *free precession* (because there is no external force acting on the composite particle). The angular velocity of this free precession is called the **Larmor frequency** ω_{L}. It is independent of the initial direction of $\boldsymbol{e}_{\text{en}}$ and given exactly by $E_{\text{HFS}}/\hbar$. Hence, the Larmor frequency of the compound spin defines the second.

The two states $\boldsymbol{e}_{\text{en}} = \boldsymbol{e}_{\uparrow\downarrow} = \boldsymbol{e}_x$ and $\boldsymbol{e}_{\text{en}} = \boldsymbol{e}_{\downarrow\uparrow} = -\boldsymbol{e}_x$ also play a particular role. In the first state, the electron spin is up and the nucleus state down. In the second state, both directions are reversed. These two states are the only states in which both the electron and the nucleus have their own spin value. All other states on the Bloch sphere, including the states with the maximum and minimum energy, $\boldsymbol{e}_{\text{en},0}$ and $-\boldsymbol{e}_{\text{en},0}$, respectively, are entangled states, which only exist in quantum theory.

Rotating frame. To ease the description of the motion of $\boldsymbol{e}_{\text{en}}$, we first multiply it with the Larmor frequency ω_{L} and write $\boldsymbol{A}_0 := \omega_{\text{L}} \boldsymbol{e}_{\text{en},0}$. Then, the Bloch vector of the compound spin rotates around $\boldsymbol{A}_0$, with the angular velocity given by the magnitude $|\boldsymbol{A}_0| = \omega_{\text{L}}$ of this vector, according to the equation[19]

[18] If, in quantum theory, two states like the spin-up and the spin-down states happen to have the same energy, they are called degenerate. With additional interactions, the energies can become different. This is called *energy splitting* and is caused, e. g., by the interaction between our two spins. Relativistic effects on the atom lead to the *fine structure* in the energy spectrum, and when the nucleus is involved, we talk about *hyperfine-structure effects*. So, hyperfine splitting is the splitting of the degeneration of the two states $\pm\boldsymbol{e}_{\text{en},0}$.

[19] This is the equation of motion for *Larmor precession*, the precession of a magnetic moment in an external magnetic field.

$$\frac{d}{dt}\boldsymbol{e}_{\text{en}}(t) = \boldsymbol{A}_0 \times \boldsymbol{e}_{\text{en}}(t).$$

Because of $\boldsymbol{e}_{\text{en}} \cdot (\boldsymbol{A}_0 \times \boldsymbol{e}_{\text{en}}) = 0$, the tip of the vector always lies on a plane perpendicular to $\boldsymbol{A}_0$.

We can use a coordinate system that rotates with the angular velocity ω around $\boldsymbol{e}_z$. In this coordinate system, the angular velocity of the Bloch vector $\boldsymbol{e}'_{\text{en}}$ is given by $\delta\omega := \omega_{\text{L}} - \omega$, and we have to use $\boldsymbol{A}'_0 = (\omega_{\text{L}} - \omega)\boldsymbol{e}'_{\text{en},0} = \delta\omega \cdot \boldsymbol{e}_{\text{en},0}$ instead of $\boldsymbol{A}_0$ (see Fig. 9.9(2b)). In this rotating coordinate system, the Bloch vector $\boldsymbol{e}'_{\text{en}}$ rotates with angular velocity $\delta\omega$. For the special case $\omega = \omega_{\text{L}}$, $\boldsymbol{A}'_0 = 0$ and the Bloch vector $\boldsymbol{e}'_{\text{en}}$ becomes static.

If you know about *nuclear magnetic resonance* (NMR), you will notice that our quantum description of the Cs compound spin corresponds exactly to the classical description of the nuclear spin in NMR, where $\boldsymbol{A}_0$ here corresponds to the static external magnetic field $\boldsymbol{B}_0$ in NMR.

9.4.5 Acting on the Compound Spin

So far, we have considered the free precession of the compound spin. But we can also **manipulate the spin** with an electromagnetic wave. The *electric field* of such an electromagnetic wave does not interact with the compound spin, because our compound particle does not have an electric dipole moment. Therefore, we can ignore the electric field of the electromagnetic wave. The *magnetic field* of the electromagnetic wave, however, does interact with the magnetic moment of our compound particle (or, if you want, independently with the magnetic moments of the electron and the nucleus).

The direction of the magnetic field in an electromagnetic wave is perpendicular to the traveling direction of said wave. In our case, we must restrict ourselves to electromagnetic waves with the magnetic field pointing in the direction of $\boldsymbol{e}_{\text{en},0}$. Magnetic fields with other directions entail that the electron and the nucleus spin no longer be parallel, and this breaks our description of the compound particle as a spin-1/2 particle.

We use a monochrome electromagnetic wave

$$\boldsymbol{B}(t) = B_0 \boldsymbol{e}_x \cos(\omega_{\text{mw}} t)$$

with a frequency ω_{mw} close to the Larmor frequency ω_{L} of the compound particle; this is an electromagnetic wave in the microwave regime (which motivates the index "mw").[20] The frequency deviation is given by $\delta\omega = \omega_{\text{L}} - \omega_{\text{mw}}$.

The effect of this microwave on the spin of the compound particle is most simply described in the frame rotating with the angular frequency ω_{mw} of the electromagnetic

[20] A microwave oven operates at 2.45 GHz. This is not very far from the 9.19 GHz of the Cs beam atomic clock.

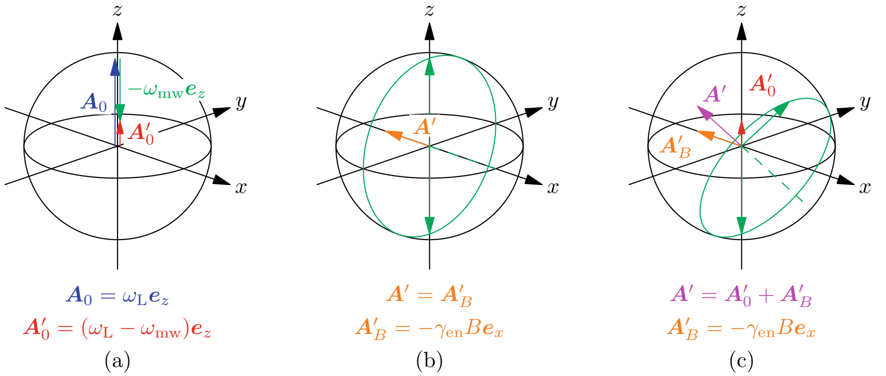

z
y
x
A_0
$-\omega_{\rm mw}e_z$
A'_0
$A_0 = \omega_{\rm L} e_z$
$A'_0 = (\omega_{\rm L} - \omega_{\rm mw}) e_z$
(a)
z
y
x
A'
$A' = A'_B$
$A'_B = -\gamma_{\rm en} B e_x$
(b)
z
y
x
A'
A'_0
A'_B
$A' = A'_0 + A'_B$
$A'_B = -\gamma_{\rm en} B e_x$
(c)

Fig. 9.10 Rotation of the Bloch vector by the magnetic field of the microwave: **a** moving from the laboratory frame with $\boldsymbol{A}_0$ to the frame rotating with angular velocity $\omega_{\rm mw}$ with $\boldsymbol{A}'_0$ (without the electromagnetic wave); **b** in the rotating frame for the resonance case; **c** in the rotating frame for the off-resonance case

wave (see Fig. 9.10a, this is still without the electromagnetic field). To include the magnetic field of the electromagnetic wave in this description, we have to add the vector $-\gamma_{\rm en}\boldsymbol{B}\boldsymbol{e}_x$ ($\gamma_{\rm en}$ is the gyromagnetic ratio of the compound particle and indicates how strongly it reacts to a magnetic field) to our vector $\boldsymbol{A}'_0$, so the Bloch vector now will rotate around the vector $\boldsymbol{A}' := \boldsymbol{A}'_0 - \gamma_{\rm en}\boldsymbol{B}\boldsymbol{e}_x$. Note that this vector does not point in the z-direction anymore, and therefore the Bloch vector $\boldsymbol{e}_{\rm en}$ does not rotate around the z-axis anymore.

In the special case when $\omega_{\rm mw} = \omega_{\rm L}$ (the microwave frequency is *in resonance* with the Larmor frequency of the compound spin), we have $\boldsymbol{A}' = -\gamma_{\rm en}\boldsymbol{B}\boldsymbol{e}_x$ and the Bloch vector rotates with the **Rabi frequency** $\omega_{\rm R} = \gamma_{\rm en}B$ around the (negative) x-axis (see Fig. 9.10b). Of practical importance are short microwave pulses that rotate the Bloch vector by a certain angle. A microwave pulse that rotates the Bloch vector by $\pi/2$, for instance, is called a $\boldsymbol{\pi/2}$**-pulse**.

If, however, the microwave frequency is not in resonance with the Larmor frequency and therefore $\delta\omega = \omega_{\rm L} - \omega_{\rm mw} \neq 0$, the vector $\boldsymbol{A}'$ no longer lies in the x-y-plane. Therefore, the Bloch vector $-\boldsymbol{e}_{\rm en,0}$ is rotated along the circle shown in Fig. 9.10c, and, in particular, cannot be rotated into $\boldsymbol{e}_{\rm en,0}$.

9.4.6 The Caesium Beam Atomic Clock

The Ramsey experiment. Now, we have all the ingredients that we need to understand how the Cs beam atomic clock works. The basis of the Cs beam atomic clock is a **Ramsey experiment**, which is very similar to an interference experiment with a photon in a *Mach-Zehnder interferometer* (MZI). In the case of an MZI, in between the first and the second beam splitters, the photon is in a *superposition* of traveling one way and traveling the other way. At the second beam splitter, these possibilities

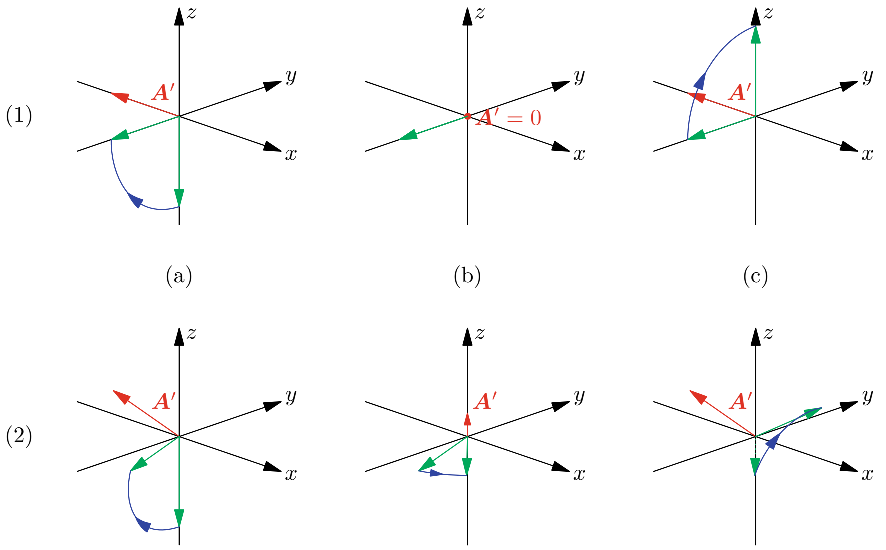


Fig. 9.11 Evolution of the compound spin of Cs in the Ramsey experiment of the Cs beam atomic clock. The coordinate system is the frame rotating with angular velocity ω. (1) In resonance; (2) off-resonance. Steps: **a** first $\pi/2$-pulse; **b** after free propagation for time T; **c** second $\pi/2$-pulse

interfere and in the case of a symmetric MZI (equal path lengths), the photon will emerge from the second beam splitter with certainty at one particular exit. In the *Ramsey experiment*, the Cs atoms cross a first microwave cavity, where their spin state is changed into the *superposition* $\boldsymbol{e}_{\text{en}} = -\boldsymbol{e}_y$ of the two states $\boldsymbol{e}_{\text{en},0}$ and $-\boldsymbol{e}_{\text{en},0}$, after which it travels freely until it crosses a second microwave cavity, where the spin is changed into $\boldsymbol{e}_{\text{en},0}$ and, if $\omega_{\text{mw}} = \omega_{\text{L}}$, a measurement yields spin up with certainty.

Let us discuss step by step what happens to a Cs atom (see Fig. 9.7). First, the Cs atom travels through the *preparation magnet*, a SG device. Suppose it leaves it at the spin-down exit. Then, its state is given by $-\boldsymbol{e}_{\text{en},0}$, as shown in Fig. 9.11(1a). The Cs atom then crosses the first microwave cavity. If the microwave frequency ω_{mw} is exactly equal to the Larmor frequency ω_{L}, the compound spin state will be rotated by exactly $\pi/2$ around the (negative) x-axis and will then point in the negative y-direction (see Fig. 9.11(1b)). After leaving the cavity, the compound spin evolves freely for some time T and, in the resonance case and the rotating frame, this means that it does not change the compound spin's direction (see Fig. 9.11(1c)). Then, the atom again crosses a microwave cavity and, in the resonance case, gets rotated again by $\pi/2$. In total, the spin gets rotated by π, and consequently then points in the direction of $\boldsymbol{e}_{\text{en},0}$ (see Fig. 9.11(1d)). When its spin in the z-direction is measured in the *analyzing magnet*, it always leaves this device at the spin-up exit.

Suppose now that the microwave frequency ω_{mw} is slightly different from the Larmor frequency. Then, in the first microwave cavity, the rotation of the Bloch vector is still close to $\pi/2$ and the Bloch vector still ends up approximately in the (negative) x-direction as long as we are close to resonance ($|\omega_{\text{mw}} - \omega_{\text{L}}| \ll \omega_{\text{R}}$), which is the relevant case. However, while the Cs atom travels, its Bloch vector

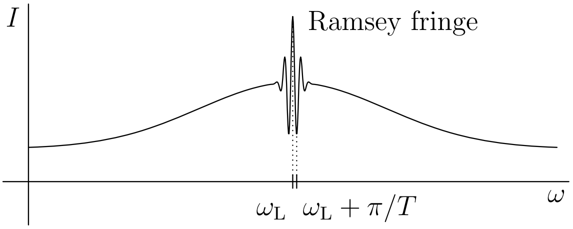


Fig. 9.12 Detector signal (relative number of spin-up measurement results) as a function of the microwave frequency ω_{mw}

rotates slowly with the angular velocity $\delta\omega = \omega_L - \omega_{\mathrm{mw}}$ around the z-axis (in the frame rotating with angular velocity ω_{mw}) and, after the time T, it is rotated by the angle $\delta\omega \cdot T$. When it crosses the second microwave cavity, the Bloch vector, in general, is not rotated in the z-direction and therefore, in the *analyzing magnet*, there will be spin-down detections. These spin-down detections indicate that the microwave frequency and the Larmor frequency (which defines the second) are not in resonance.

The *measured signal* (or the probability of detecting spin-up) as a function of the microwave frequency ω is shown in Fig. 9.12. For $\omega_{\mathrm{mw}} = \omega_L$, the signal is maximal; around $\omega = \omega_L$, we see an interference pattern, and the peaks are called *Ramsey fringes*. What the *feedback electronics* now does is to continuously vary the microwave frequency a bit around ω_L and to adapt the microwave frequency such that the detector signal is kept maximal. In this way, the microwave frequency ω_{mw} is kept equal to the Larmor frequency ω_L and the microwave cycles can be counted and divided by 9.192.631.770 such that the resulting signal can be used to drive the clock.

The microwave frequency is typically generated by a *voltage-controlled crystal oscillator* (VCXO) (see Fig. 9.7), which can be a highly temperature-stabilized quartz crystal whose frequency can be changed in a small interval by applying a voltage across it.

Precision of the Cs beam atomic clock. How stable (precise) and accurate is a Cs beam atomic clock? The **stability** of the clock is mainly given by the precision of the measurement of the deviation $\delta\omega$ of the microwave frequency from the Larmor frequency. To get a good stability, one needs a narrow central Ramsey fringe. The width of a Ramsey fringe is given by π divided by the flight time T of the atoms. This width must be as small as possible, therefore, one desires long flight times. The average velocity of the atoms is given by their temperature, and typically is on the order of magnitude of 100 m/s, and the separation of the Ramsey zones is about 1 m. Therefore, $T \sim 100$ ms, which gives a linewidth of about $\pi/T \approx 50$ Hz. To reach a fractional uncertainty of about 10^{14} (as in the case of the CS2), one needs to resolve the fringe to one part in a million or to 5×10^{-5} Hz. This gives a stability of 5×10^{-5} Hz/9.2 GHz $\approx 10^{-14}$ and requires a signal-to-noise ratio on the order of magnitude of 10^6, which is only possible if the flux of atoms is pretty large. A higher flow of atoms, however, makes it more likely that there will be intra-beam collisions, which are a source of errors and limit the separation of the Ramsey zones,

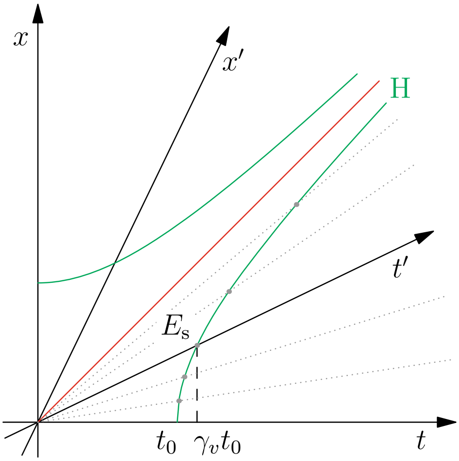


Fig. 9.13 Scaling of the axes of Bob

and therefore of T. This limits the precision of the measurement or the stability of the clock.

There are several effects that limit the clock's **accuracy**. Among these are the influence of magnetic fields on the Larmor frequency of the hyperfine-split ground state, the (relativistic) Doppler effect, which also changes the resonance frequency because of time dilation of the moving atoms, and several more.

9.5 The Spacetime Diagram III: Scales

The scaling in spacetime transformations. Now, we continue with the discussion on time dilation that we started in Sects. 7.2 and 7.6. There, we introduced the spacetime diagram and learned how Alice, in her spacetime diagram, has to draw the space and time axes of Bob. The question on the scale, however, is still open: where on Bob's axes does Alice have to put the length "1 m" and the time "1 s"?

As always, we take Bob to move with the velocity v relative to Alice, on the trajectory $x = vt$.

We start with the scale of the t'-axis (see Fig. 9.13) and ask where on the t'-axis Bob's "1 s"-tick is, i. e., where on the t'-axis is the event E_s at which Bob's clock shows $t'_s = 1$ s? Let $E_0 = (t_0, 0)$ denote the event at which Alice's clock shows 1 s. Due to time dilation, the event E_s is simultaneous for Alice with her event with $t = \gamma_v t_0$ on the t-axis. So, we mark t_0 and $t_s = \gamma_v t_0$ on the t-axis and, by drawing a line of simultaneity for Alice, we locate the event E_s, which, for Alice, has the coordinates $t_s = \gamma_v t_0$ and $x_s = vt_s$, or $E_s = (\gamma_v t_0, v\gamma_v t_0)$.

Now, we imagine "many Bobs", all inertial observers, who fly by Alice with different relative velocities v and follow the trajectories $x = vt$, respectively. The events E_s of these "Bobs" (the events obviously depend on v) yield a curve H in the spacetime diagram. By plugging in the coordinates of the events E_s, one sees easily that this curve is given by

$$c^2t^2 - x^2 = c^2t_0^2.$$

This curve is a hyperbola, which is symmetrical to Alice's time axis and intersects this axis in the event $(t = t_0, x = 0)$.

Exercise 35: The hyperbola in Fig. 9.13 intersects the t-axis perpendicularly, i. e., parallel to the x-axis. Show that the hyperbola also intersects the t'-axis parallel to the x'-axis.

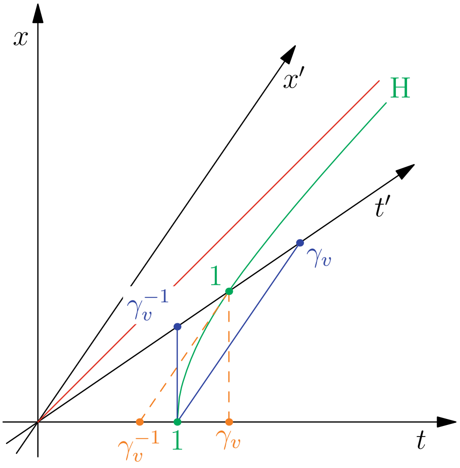
The question on the scale of the t'-axis has now been answered. To find out where "1 s" is located on the t'-axis, we draw a hyperbola $x(t) = \pm c\sqrt{t^2 - t_0^2}$, which is located symmetrical to Alice's time axis and intersects it at $t_0 = 1$ s. This hyperbola then intersects Bob's time axis where 1 s has passed for Bob.

The scale of the space axis is determined analogously: to find out where "1 m" is on the x'-axis, we draw a hyperbola $x(t) = +\sqrt{x_0^2 + c^2t^2}$, symmetrically to Alice's x-axis, such that it intersects the axis at $x_0 = 1$ m. The hyperbola then intersects Bob's x'-axis where "1 m" is for Bob. To prove this, one applies the length contraction in the same way as the time dilation above.

Scale transfer: The hyperbola $c^2t^2 - x^2 = c^2t_0^2$ intersects the t-axis at $t = t_0$ and the t'-axis at $t' = t_0$ and is used to transfer the time scale from Alice's to Bob's coordinate system.

The hyperbola $x^2 - c^2t^2 = x_0^2$ intersects the x-axis at $x = x_0$ and the x'-axis at $x' = t_0$ and is used to transfer the time scale from Alice's to Bob's coordinate system.

Fig. 9.14 Procedure for transferring units from one inertial reference frame to another



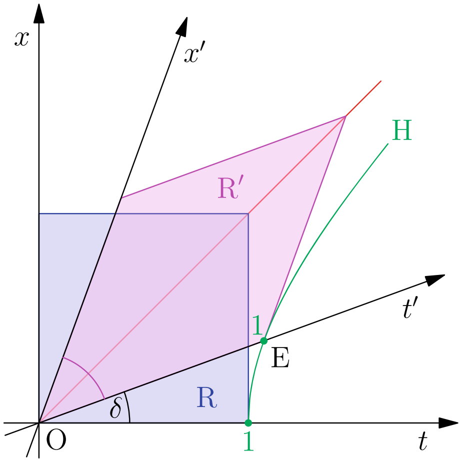


Fig. 9.15 Regarding Exercise 36. One corner of the square R (the rhombus R′) lies on the intersection of the t-axis (t'-axis) with the hyperbola $c^2t^2 - x^2 = 1$

To determine rapidly the units of Bob's coordinate system in a drawing, one can use the construction in Fig. 9.14 (remember that $\gamma^{-1} < 1 < \gamma$ for $v \neq 0$):

- One starts on the t-axis at $t = 1$ and draws a line parallel to the x-axis; the latter intersects the t'-axis at $t' = \gamma^{-1}$.
- One draws a line parallel to the x'-axis and ends up at intersecting the t'-axis at $t = \gamma$.
- The likewise holds for Bob. Starting on the x'-axis at $t' = 1$, one draws a line parallel to the x'-axis. This line intersects the t-axis at $t = \gamma_v^{-1}$. And if one draws a line parallel to the x-axis, one ends up at $t = \gamma_v$.

Exercise 36: Calculate the area of the rhombus R′ (see Fig. 9.15) in Alice's measurement units:

1. Determine the t-coordinate of the event E with $(t' = 1, x' = 0)$. (e. g., by using the hyperbola $c^2t^2 - x^2 = 1$).
2. Calculate from the t-coordinate above the length l of the line segment $\overline{OE}$ (in Alice's coordinates!).
3. The area of a rhombus is given by $A = l^2 \sin\alpha$, where l is the length of an edge and α one of the inner angles. Express the angle α via the angle δ, and this one in turn by v/c.
4. Now, the formula $\cos(2\delta) = (1 - \tan^2\delta)/(1 + \tan^2\delta)$ could be helpful.

The result shows that a rhombus with two edges coinciding with Bob's coordinate axes and that, in Bob's coordinate system, has an edge length of l (as given above) has the area 1 in Alice's coordinate system. In this way, Alice can determine the length and time units of Bob (as an alternative to the hyperbola method).

The spacetime distance. In Euclidean geometry, in the two-dimensional plane, the distance $d = \sqrt{(\Delta x)^2 + (\Delta y)^2}$ between two points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$,

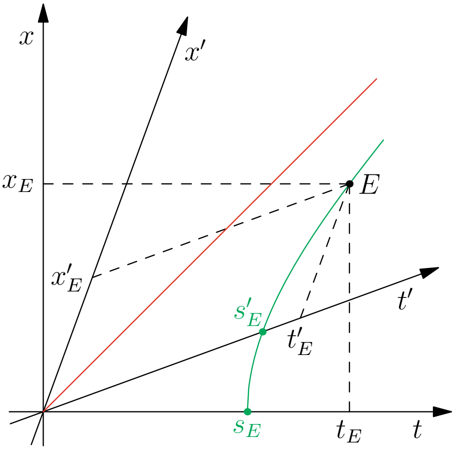


Fig. 9.16 Proof that $s^2 = (ct)^2 - x^2$ is invariant

where $\Delta x = x_2 - x_1$ and $\Delta y = y_2 - y_1$, is independent of the orientation of the coordinate system (and how it is rotated). The distance is an *invariant with respect to rotations*. In the same way, the quantity

$$s = \sqrt{(c\Delta t)^2 - (\Delta x)^2} \, ,$$

which is associated with two events $E = (t_E, x_E)$ and $F = (t_F, x_F)$ in spacetime, is invariant when changing from one inertial frame to another. The invariant s is the **spacetime distance** of two events.[21] We now proof the invariance of the spacetime distance.

To do this, we have drawn a hyperbola $(ct)^2 - x^2 = \text{const.}$ into Fig. 9.16. This hyperbola passes through the event E, which, in Alice's coordinate system, has the coordinates (t_E, x_E). On the x-axis, we have $x = 0$, and therefore, the hyperbola intersects the t-axis at

$$s_E = \sqrt{(ct_E)^2 - x_E^2}.$$

From Bob's point of view, the curve is also a hyperbola, which is symmetric to his t'-axis and passes through E. For him, the event E has the coordinates (t'_E, x'_E). The hyperbola intersects his t'-axis at

$$s'_E = \sqrt{(ct'_E)^2 - x'^2_E}.$$

[21] The change from one inertial frame to another is called a Lorentz transformation (see Chap. 11). The *Euclidean distance d* of two points in the two-dimensional plane has the same relation to the *rotation* as the *spacetime distance s* of two events in spacetime to the *Lorentz transformation*.

Because the hyperbola defines the scale of the axis, we have $s_E = s'_E$, and therefore

$$(ct_E)^2 - x_E^2 = (ct'_E)^2 - x'^2_E \tag{9.2}$$

for the coordinates of an event E from the point of view of two arbitrary inertial observers.

This is also valid for the coordinate differences. Take an additional event F with the coordinates (t_F, x_F) in Alice's and (t'_F, x'_F) in Bob's coordinate system. Now we move the coordinate system such that the event F is the new origin and apply (9.2). Then, with the abbreviations $\Delta t = t_E - t_F$, $\Delta x = x_E - x_F$, $\Delta t' = t'_E - t'_F$ and $\Delta x' = x'_E - x'_F$, it follows that

$$(c\Delta t)^2 - (\Delta x)^2 = (c\Delta t')^2 - (\Delta x')^2.$$

The spacetime distance between two events is the same for all inertial observers. This is a very important relation, and we will show later that length contraction and time dilation can be derived directly from it. With differentials, we can write it as

$$ds^2 = c^2\,dt^2 - dx^2. \tag{9.3}$$

9.6 The Relativistic Doppler Effect

9.6.1 Longitudinal Doppler Effect

We discussed the classical Doppler effect for the special case of one space dimension in considerable detail in Sect. 4.2.4 and got the general result (4.9) for the change of the frequency. We already pointed out there that it is not the relative velocity of source and observer that appears in the formula, but the velocities of both the source and the observer, relative to the medium. This is clearly not possible for light waves, as there is no medium. Equation (4.9), however, has one more defect. The derivation used the Galilean addition of velocities, which, for large velocities, is no longer correct. Another way to say this is through time dilation, which causes the clock of the source to move slowly from the point of view of the observer. These considerations must have consequences on the Doppler effect. As a remedy, we investigate the **relativistic Doppler effect**.

Longitudinal Doppler effect. The situation that we will discuss now is outlined in Fig. 9.17. Bob moves relative to Alice with the velocity $v > 0$ in the positive x-direction and carries the source, whereas Alice is the observer. Now let the source emit a wave with a well-defined frequency and have the node with phase $\varphi = 0$ be at the common origin of Alice's and Bob's coordinate system. Let the node with phase $\varphi = 2\pi$ of the wave be located at the event E_S.

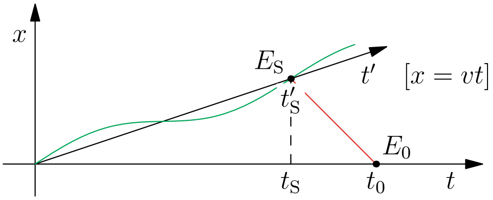


Fig. 9.17 Derivation of the relativistic formula for the longitudinal Doppler effect

For Bob (or the source, respectively), the period of the emitted signal is given by the time elapsed between the origin and the event E_S. This is exactly the t'-coordinate t'_S of E_S. The second node moves with the speed of light from event E_S to Alice and arrives there at event E_0. For Alice, the period of the light wave is therefore given by the time elapsed between the origin and the event E_0, which is t_0.

To get hold of the relation between t'_S and t_0, we first calculate the time t_S at which the event E_S happens for Alice. Then, we are only left with finding the trajectory of the node traveling from E_S to E_0.

Because of time dilation, $t_S = \gamma_v t'_S$. Thus, the node's trajectory is given by $x(t) = -c(t - t_S) + vt_S$. This intersects the t-axis at $t_0 = (1 + v/c)t_S = \gamma_v \cdot (1 + v/c)t'_S$. The ratio between the frequencies ν_S of the source and ν_O of the observer is equal to the inverse of the ratio of the periods, $\nu_O/\nu_S = t'_S/t_0$. Therefore, the formula for the **longitudinal Doppler effect** is

$$\frac{\nu_O}{\nu_S} = \frac{1}{\gamma_v \cdot (1 + v/c)} = \sqrt{\frac{1 - v/c}{1 + v/c}}. \tag{9.4}$$

This is called the *longitudinal* Doppler effect, as it describes the case when the source moves radially away ($v > 0$) or toward ($v < 0$) the observer.[22]

Note that $1 - (v/c)^2 = (1 - v)(1 + v)$, and therefore the product of γ_v and $1 + v/c$ is equal to $\sqrt{(1 + v/c)/(1 - v/c)}$. For the source moving away from the observer, $v > 0$ and $\nu_O < \nu_S$. There is a frequency shift to lower frequencies.

In Exercise 37, we show how this becomes equal to the classical Eq. (4.9) in the case of small velocities.

Exercise 37: Show that Eq. (9.4) for the relativistic Doppler effect for small velocities ($v/c \ll 1$) yields the same results as Formula (4.9) for the classical Doppler effect. You will need the following approximation, valid for $x \ll 1$: $1/(1 + x) \approx 1 - x$.

[22] "Longitudinal" means "in the lengthwise direction".

Exercise 38: We performed the derivation of (9.4) from the perspective of the inertial frame where the observer is at rest. Show that any inertial frame could have been used. To do so, assume that Einstein is at rest in an inertial frame and restrict oneself to one dimension where Einstein is located at the origin. Alice, who holds the source, is located at $x_A > 0$ and moves relative to Einstein with the velocity $u > 0$ away from him. Bob, the observer, is located at $x_B < 0$ and moves relative to Einstein with the velocity $w > 0$ toward him. The source radiates with the frequency ν_S and the observer measures the frequency ν_O.

Show the following:

1. For Einstein, the source radiates with the frequency

$$\nu_E = \nu_S \sqrt{\frac{1 - u/c}{1 + u/c}}.$$

2. Bob could consider Einstein, who moves with velocity $-w$ relative to Bob, as the source. Then, he would measure the frequency

$$\nu_O = \nu_E \sqrt{\frac{1 + w/c}{1 - w/c}}.$$

3. Now show that, by combining these two formulas, Bob gets

$$\nu_O = \nu_S \sqrt{\frac{1 - v/c}{1 + v/c}},$$

where

$$v = u \ominus w = \frac{u - w}{1 - uw/c^2}.$$

9.6.2 Transversal Doppler Effect and the General Formula

Retarded location and velocity of the source. The longitudinal Doppler effect, with its radial (or longitudinal) motion, is always in one space dimension. In two (or three) dimensions, the situation with the Doppler effect becomes considerably more difficult. Due to retardation, the wave that the observer measures was emitted by the source some time ago, when the source had a different location and, possibly, velocity. Therefore, for the Doppler effect, the direction $\boldsymbol{e}$ of the source and its velocity $\boldsymbol{v}$ at the retarded time (i.e., when the wave was emitted) is relevant. Both the direction and the velocity of the source are relative to the observer. The direction of the source

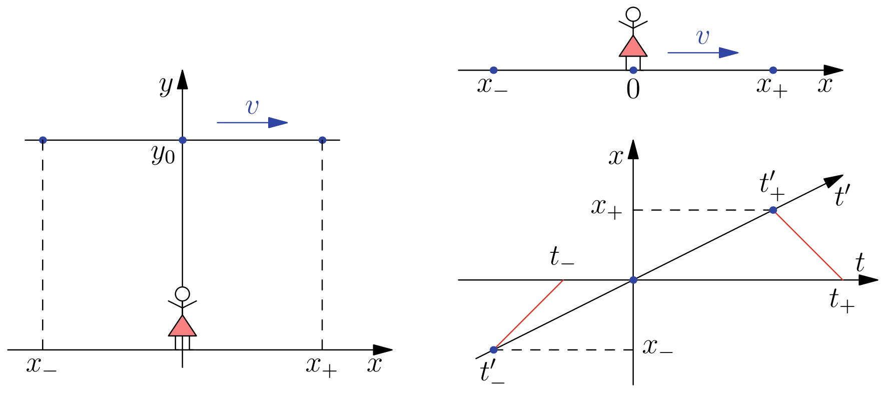


Fig. 9.18 Derivation of the relativistic formula for the transversal Doppler effect

at the retarded time is parallel to the wave vector at the location of the observer. The notion of simultaneity here is that of the observer.

Transversal Doppler effect. As we mentioned already, (9.4) describes the *longitudinal Doppler effect*. What happens when the source passes the observer?

Suppose that, as shown in Fig. 9.18 on the left, the source moves on the trajectory $x(t) = vt$, $y(t) = y_0$ and the observer is located at the coordinate system's origin. Then, the frequency change of a wave becomes dependent on the position of the source when the wave was emitted.

To find the general formula, we start with another special case, the case when the velocity of the source is exactly perpendicular to the line passing through the source and the observer.[23] This is called the **transversal Doppler effect**. Then, suppose that the emitted wave had the phase $\varphi = -\pi, 0, \pi$ when the source was at $x = x_+ = -x_0$, $x = 0$, $x = x_- = x_0$, respectively. We have to determine the times when the wavefronts corresponding to the nodes $\varphi = -\pi, +\pi$ arrive at the observer's location. As long as both wavefronts have to travel the same distance, the coordinate y_0 does not matter (which is why we have chosen them to be at $x_+ = -x_-$). We can even choose $y_0 = 0$ and get the situation shown in Fig. 9.18 on the right. The nodes are emitted at times $t'_- = -T'/2$ and $t'_+ = +T'/2$ in the reference frame of the source. The arrival times at the observer's location can now be calculated with the formula for the longitudinal Doppler effect. We get

$$t_\pm = \pm\sqrt{\frac{1 \pm \beta}{1 \mp \beta}}\frac{T'}{2}.$$

The difference $T = t_+ - t_-$, which is the wave period for the observer, becomes

[23] As pointed out, this means the retarded position and velocity.

$$T = \frac{1}{2}\left(\sqrt{\frac{1-\beta}{1+\beta}} + \sqrt{\frac{1+\beta}{1-\beta}}\right) T' = \gamma T'.$$

Taking into consideration $\nu_{\mathrm{O}}/\nu_{\mathrm{S}} = T'/T$, for the *transversal Doppler effect*, we get

$$\frac{\nu_{\mathrm{O}}}{\nu_{\mathrm{S}}} = \frac{1}{\gamma_v}. \tag{9.5}$$

For the observer, the frequency of the source is smaller than it actually is. Note that, classically, there is no transversal Doppler effect.

General formula for the Doppler effect. Comparing the Formula (9.4) for the longitudinal and Formula (9.5) for the transversal Doppler effect, we recognize that, in both cases, the time dilation of the moving source contributes to the Doppler effect. For the longitudinal Doppler effect, we have an additional contribution from the fact that the source moves toward the observer or away from it and the changing traveling time for the signal has to be considered. This shows that, in the general case, we must take into account the time dilation and the *radial component* of the source's velocity $\boldsymbol{v}$, because this determines the change of the traveling time. If $\boldsymbol{e}$ is the unit vector pointing from the observer to the source, this radial component v_{r} of the source's velocity is given by $v_{\mathrm{r}} = \boldsymbol{e}\boldsymbol{v}$. Therefore, in general, for the *relativistic Doppler effect*, we have

$$\frac{\nu_{\mathrm{O}}}{\nu_{\mathrm{S}}} = \frac{1}{\gamma_v \cdot (1 + \boldsymbol{e}\boldsymbol{v}/c)}. \tag{9.6}$$

Note that while, in the factor $1 + \boldsymbol{e}\boldsymbol{v}$, only the radial component of the source's velocity appears, the γ-factor still features the magnitude of the full velocity of the source. For the time dilation, it does not matter in which direction the source moves.

9.7 The Experiment by Ives and Stilwell

For the *transversal Doppler effect*, the ratio between the observed frequency ν_{O} and the frequency ν_{S} of the signal as seen by the source, according to (9.5), is given directly by the γ-factor and, as we stated, the reason is *time dilation* for the moving source. This gives a direct way to measure time dilation, and the Americans Ives and Stilwell, in 1938, were the first to perform this important experiment [IvesStilwell38].

The basic idea is to use an atom as a source. Atoms in excited states emit light at well-konwn discrete frequencies that correspond to transitions from one state to another and can be seen as a clock. If the atoms move, the observer measures frequencies that are modified according to the Doppler effect.

Ives and Stilwell used a discharge tube filled with hydrogen gas at a very low pressure (see Fig. 9.19). Hydrogen gas consists of hydrogen molecules, most are diatomic hydrogen H_2, some triatomic hydrogen H_3. By natural processes (like

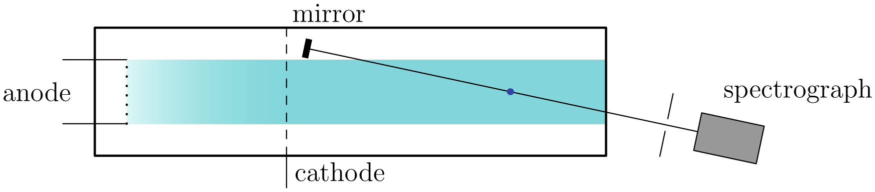


Fig. 9.19 Setup of the Ives-Stilwell experiment. For the description, see the text

radioactivity), there are always some of these molecules that are ionized. A strong electric field between an anode and a cathode now brings ions and electrons separately and accelerates them to high velocities. Some ions H_2^+ and H_3^+ fly through the perforated cathode and travel with constant velocity in the tube (the turquoise beam in the figure). At some point, they catch a free electron and recombine by emitting light of well-known frequencies (Balmer lines) in all directions (the blue dot in the figure). A part of the radiation directly reaches the detector, which sees the emitting hydrogen molecule as approaching it, while another part arrives at the mirror, gets reflected and then reaches the detector, which sees the emitting hydrogen molecule as receding from it. Due to the longitudinal Doppler effect, these rays experience different frequency shifts.

This experiment is very challenging and the reason is the following. If ϑ is the angle between the direction $\boldsymbol{e}$ of the source and the velocity[24] $\boldsymbol{v}$ of the source relative to the observer, we have $\boldsymbol{ev} = v \cos\vartheta$ and, from (9.6), we get[25]

$$\frac{\lambda_\mathrm{O}}{\lambda_\mathrm{S}} = \gamma_v \cdot (1 + \beta \cos\vartheta). \tag{9.7}$$

For $\vartheta = \pi/2$, this is the transversal Doppler effect, which, because of $\gamma_v \approx 1 - \beta^2/2$, is an effect that is quadratic in β, a *second-order effect*. Due to the fact that β in experiments is usually very small, a second-order effect is minuscule.

If ϑ is not exactly $\pi/2$, then the term linear in β in the second factor on the right side of (9.7) does not vanish. This linear term comes from the longitudinal Doppler effect, which is a *first-order effect* and is much larger than a second-order effect.

Let $\vartheta = \pi/2 + \delta$, with a small angle δ. Then,

$$\frac{\lambda_\mathrm{O}}{\lambda_\mathrm{S}} = \gamma_v \cdot (1 - \beta \sin\delta) \approx (1 + \beta^2/2)(1 - \beta\delta) \approx 1 - \beta\delta + \beta^2/2. \tag{9.8}$$

As soon as δ becomes comparable to β, the first-order longitudinal Doppler effect completely wipes out the second-order transversal Doppler effect and it is no longer

[24] Both the direction $\boldsymbol{e}$ and the velocity $\boldsymbol{v}$ of the source are meant in the moment when the source emitted the wave, where the notion of simultaneity is that of the observer.

[25] We use wavelengths instead of frequencies because the formulas become considerably easier and because, in the experiment, wavelengths are measured, not frequencies.

the time dilation that is measured. In the experiment by Ives and Stilwell, β was several times smaller than $1/100$, so a deviation of only 0.5° from the direction $\vartheta = \pi/2$ would already completely ruin the experiment.

The ingenious idea of Ives and Stilwell was to measure in two opposite directions by using a mirror and to add the two wavelengths (see Fig. 9.19). Take the directions ϑ_1 and $\vartheta_2 = \vartheta_1 + \pi$, then, the two measured wavelengths are

$$\lambda_{\mathrm{O},1} = \lambda_{\mathrm{S}}\gamma_v \cdot (1 + \beta \cos \vartheta_1) \quad \text{and} \quad \lambda_{\mathrm{O},2} = \lambda_{\mathrm{S}}\gamma_v \cdot (1 - \beta \cos \vartheta_1)$$

and

$$\frac{\lambda_{\mathrm{O},1} + \lambda_{\mathrm{O},2}}{\lambda_{\mathrm{S}}} = 2\gamma_v. \tag{9.9}$$

If there is a small deviation and we have $\vartheta_2 = \vartheta_1 + \pi + \delta$, we get, instead of $\cos \vartheta_1 + \cos \vartheta_2$,

$$\begin{aligned} \cos \vartheta_1 + \cos \vartheta_2 &= \cos \vartheta_1 - [\cos \vartheta_1 \cos \delta - \sin \vartheta_1 \sin \delta] \\ &= \cos \vartheta_1 (1 - \cos \delta) + \sin \vartheta_1 \sin \delta \\ &\approx \cos \vartheta_1 (\delta^2/2) + \sin \vartheta_1 \delta. \end{aligned}$$

In the ideal case, we would choose $\vartheta_1 = 0$, because the larger linear term then vanishes. This is not possible in the Ives-Stilwell experiment, although they did choose ϑ_1 as small as possible.

In total, one gets

$$\begin{aligned} \frac{\lambda_{\mathrm{O},1} + \lambda_{\mathrm{O},2}}{\lambda_{\mathrm{S}}} &= \gamma_v \left(2 + \beta \cdot (\cos \vartheta_1 \cdot (\delta^2/2) + \sin \vartheta_1 \cdot \delta)\right) \\ &\approx 2 + \beta \cdot (\cos \vartheta_1 \cdot (\delta^2/2) + \sin \vartheta_1 \cdot \delta) + \beta^2/2 , \end{aligned}$$

which is a lot better than the direct transversal measurement. Ives and Stilwell had $\vartheta_1 = 7° \approx 0.12$, which gives us $\sin \vartheta_1 \approx 0.12$. Therefore, the measurement method chosen by Ives and Stilwell is almost ten times less susceptible to errors in the angle than the direct transversal measurement.

A helpful additional property is that, if we take the difference of the two measured wavelengths, we get

$$\frac{\lambda_{\mathrm{O},1} - \lambda_{\mathrm{O},2}}{\lambda_{\mathrm{S}}} = 2\gamma_v \cdot \beta \cos \vartheta_1 \approx 2\beta , \tag{9.10}$$

which allows us to determinate the velocity of the atoms.

Ives and Stilwell measured $\Delta\lambda$ and $\Delta'\lambda$, defined by

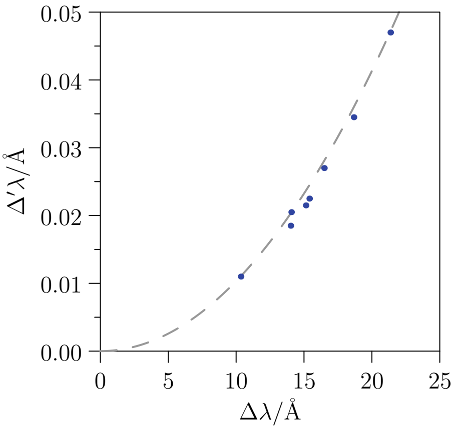


Fig. 9.20 Result of the Ives-Stilwell experiment. The blue dots are the measured values and the grey dashed curve the expectation from theory

$$\Delta'\lambda = \frac{(\lambda_{O,1} - \lambda_S) + (\lambda_{O,2} - \lambda_S)}{2} \,,$$
$$\Delta\lambda = \frac{\lambda_{O,1} - \lambda_{O,2}}{2} \,,$$

which, according to (9.9) and (9.10), gives us

$$\Delta'\lambda = \lambda_S(\gamma_v - 1) \approx \frac{1}{2}\lambda_S\beta^2 \,,$$
$$\Delta\lambda = \lambda_S\beta\gamma \approx \beta\lambda_S \,,$$

and therefore

$$\Delta'\lambda = \frac{(\Delta\lambda)^2}{2\lambda_S}.$$

By plotting $\Delta'\lambda$ over $\Delta\lambda$, they got the graph in Fig. 9.20, which is a nice confirmation of time dilation.

The Ives-Stilwell experiment confirms the formulas for the transversal relativistic Doppler effect and the time dilation.

9.8 The Experiment by Kennedy and Thorndike

In Sect. 5.2, we discussed the Michelson-Morley experiment and concluded that it can be explained with length contraction alone (as FitzGerald and Lorentz did) and that no time dilation is needed.

A relatively small modification of the experiment makes it more impactful, and this is what the Americans Kennedy and Thorndike did in 1932 [KennedyThorndike32] in the **Kennedy-Thorndike experiment**. They used an interferometer similar to that of Michelson and Morley, but with different lengths for the two interferometer arms.[26] To see what happens, we repeat the discussion in Sect. 5.2.1 with different arm lengths.

Let L_1 and L_2 be the lengths of the first and second interferometer arms, respectively. Let us further consider two orientations. In orientation 1, the motion of the interferometer relative to the supposed luminiferous aether is parallel to arm 1 while in orientation 2, the motion is parallel to arm 2.

In orientation 1, the traveling times are

$$\begin{aligned} T_{1,1} &= \frac{2L_1}{c}\frac{1}{1-\beta^2} , \\ T_{2,1} &= \frac{2L_2}{c}\frac{1}{\sqrt{1-\beta^2}} , \end{aligned}$$

and in orientation 2, we have

$$\begin{aligned} T_{1,2} &= \frac{2L_1}{c}\frac{1}{\sqrt{1-\beta^2}} , \\ T_{2,2} &= \frac{2L_2}{c}\frac{1}{1-\beta^2} . \end{aligned}$$

This leads to the following travel time differences when we rotate the apparatus by 90°:

$$\begin{aligned} \Delta T_1 = T_{1,1} - T_{2,1} &= \frac{2}{c}\left(\frac{L_1}{1-\beta^2} - \frac{L_2}{\sqrt{1-\beta^2}}\right) , \\ \Delta T_2 = T_{1,2} - T_{2,2} &= \frac{2}{c}\left(\frac{L_1}{\sqrt{1-\beta^2}} - \frac{L_2}{1-\beta^2}\right) . \end{aligned}$$

In the Michelson-Morley experiment, the interference pattern does not change if one changes the interferometer's orientation, i. e., $\Delta T_1 = \Delta T_2$. This is the case, if one assumes that there is length contraction in the traveling direction (Lorentz-FitzGerald length contraction), i. e., in the traveling times $T_{1,1}$ and $T_{2,2}$, where the length contraction leads to replacing the denominator $1-\beta$ with $\sqrt{1-\beta^2}$. With length contraction, i. e., $L_{1,0} = L_1/\sqrt{1-\beta^2}$ and $L_{2,0} = L_2$ in orientation 1 and $L_{1,0} = L_1$ and $L_{2,0} = L_2/\sqrt{1-\beta^2}$ in orientation 2, we have

[26] The challenge with interferometer arms of different lengths is that wave trains emitted at different times from the light source must be able to interfere. This is only possible if the time difference is shorter than the coherence time of the light. For the emission radiation used from the mercury line at 546.1 nm, this corresponds to a maximum interferometer arm length difference of about 16 cm.

$$\Delta T_1 = \frac{2}{c}\frac{L_{1,0} - L_{2,0}}{\sqrt{1-\beta^2}} = \Delta T_2.$$

In the case of the Michelson-Morley experiment, this vanishes, and the Lorentz-FitzGerald length contraction explains the absence of changes in the interference pattern when the measurement apparatus' orientation is changed.[27] Even if *the lengths are not equal* (although we still assume FitzGerald's length contraction), the interference pattern is the same for orientation 1 and orientation 2. The interference pattern does not depend on the direction.

Kennedy and Thorndike did not rotate their interferometer, it was firmly mounted on the floor, and they consequently only measured one traveling time difference (and, hence, interference pattern)[28]

$$\Delta T = \frac{2}{c}\frac{\Delta L}{\sqrt{1-\beta^2}}.$$

The key point is that, for different arm lengths $\Delta L = L_{1,0} - L_{2,0} \neq 0$ (with Lorentz-FitzGerald length contraction taken into account), this traveling time difference depends via γ_v on the magnitude of the velocity of the apparatus relative to the supposed aether. Kennedy and Thorndike observed the interference pattern for a very long time, more than a year. Assuming that the Sun moves with a certain velocity in the aether, due to its orbital motion around the Sun, the Earth's velocity relative to the aether must change by twice the Earth's orbital velocity in the course of a year. Kennedy and Thorndike have shown that such a motion would be detectable with their apparatus, but they did not observe any change in the interference pattern. Therefore, ΔT does not depend on v and the Lorentz-FitzGerald length contraction, which is able to explain the result of the Michelson-Morley experiment in the context of an aether theory, cannot accomodate the result of the Kennedy-Thorndike experiment.

The only concept that can rescue the aether theory at this stage is an *ad hoc* introduced time dilation $\Delta T_0 = \Delta T/\sqrt{1-\beta^2}$, which yields $\Delta T_0 = 2\Delta L/c$ for the traveling time difference in the interferometer. This time dilation was introduced by Lorentz into the aether theory, although for other reasons and many years before the Kennedy-Thorndike experiment.

[27] The Lorentz-FitzGerald contraction, however, is *ad hoc*. It "repairs" the problem, but does not really explain what goes on.

[28] If one takes the result of the Michelson-Morley experiment for granted, i.e., the speed of light does not depend on the direction, then one sees that the two interferometer arms in the Kennedy-Thorndike experiment do not have to be perpendicular to each other. They could even be parallel.

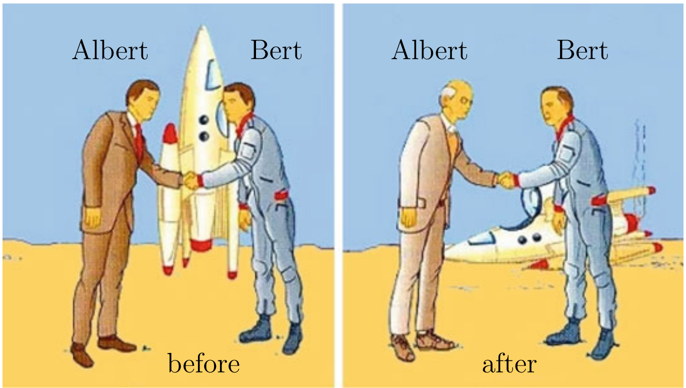


Fig. 9.21 The twins Albert (standing in here for Alice) and Bert (standing in here for Bob) just before the beginning and just after the end of Bert's space travel. Bert has aged more slowly (has stayed younger) than Albert, who remained in an inertial frame

9.9 Twin Paradox

As long as Alice and Bob move away from each other, time dilation cannot demonstrate how spectacular it is, because Alice and Bob can no longer meet to compare their clocks on site. But what happens when Bob, at some point, reverses his traveling direction and comes back to Alice? Exactly what you thought: when, after the return, both stand next to each other, they will discover that, for Bob, less time has passed than for Alice. Bob will have aged less! This effect is called the **twin paradox** (see Fig. 9.21).

The reason for this can easily be seen in Fig. 9.22. Albert[29] is an inertial observer. In the event O, he meets with his twin Bert and both synchronize their clocks. Then, Bert travels away from Albert with the constant velocity v and on the trajectory $x = vt$ or $x' = 0$ (as usual, we denote Bert's coordinates by x' and t'). At event P, Bert suddenly changes his traveling direction and, after traveling back, meets Albert again at event Q. By changing his traveling direction, Bert changed his inertial frame. For reasons of convenience, we use a new coordinate system for Bert and denote his coordinates by x'' and t''. The origin of this coordinate system coincides with Q.

Now, from O to E_1, the same time passes for Albert as for Bert on his journey from O to P. And from E_2 to Q, the same time passes for Albert as for Bert for his journey from P to Q, back to Albert. Both facts can be easily seen with our hyperbola construction.

So, the time between their encounters is longer for Albert than it is for Bert. The difference is exactly the time between E_1 and E_2. Let t_1 and t_2 be the time at which the events E_1 and E_2 occur, respectively. Then, by the time $\Delta t = t_2 - t_1$, Albert

[29] In the discussion on the twin paradox, we replace Alice with Albert and Bob with Bert so as to have a clearer situation with twins of the same gender.

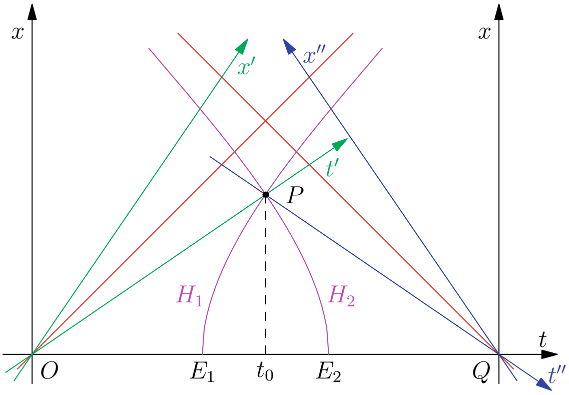


Fig. 9.22 To the twin paradox. See the text

will be older than Bert when they meet again. If you move the event P in Fig. 9.22 upward, Bert travels faster and E_1 moves closer to O. Furthermore, E_2 moves closer to Q and Δt, the age difference, becomes larger. In Exercise 39, we will determine the age difference.

In our model, Bert reverses his motion suddenly. In reality, this is, of course, not possible, because the acceleration would need to be infinite. But as an idealization, we can assume it. In Sect. 9.10, we will show what happens if Bert slowly reverses his motion.

Now, you may think: ok, Bert's clock runs more slowly than Albert's. But since Bert ages more slowly, this is not possible. Nevertheless, this is the case. All processes (including aging processes) in our body are based on similar physical principles as the clock. If the clock runs more slowly, atomic processes are also more slowly. And if atomic processes are more slowly, a human will age more slowly. Otherwise, we would be in contradiction with the principle of relativity, because certain inertial frames would be "more special" than the others.

One could arrive at the argument that Bert also could say that Albert has aged more slowly on his journey. This conclusion, however, is not tenable. In comparison to the situation when discussing length contraction and time dilation, Albert here is always in an inertial frame. Bert, however, is not. At some point, he changes his direction of motion and, indeed, feels it. For this reason, the principle of relativity is not applicable for his point of view.

9.10 Digression: Proper Time

Elaborating on the twin paradox, we now deal with the general case in which Bob moves on an arbitrary trajectory $x_B(t)$ relative to the inertial observer Alice[30] (see

[30] We don't need twins anymore.

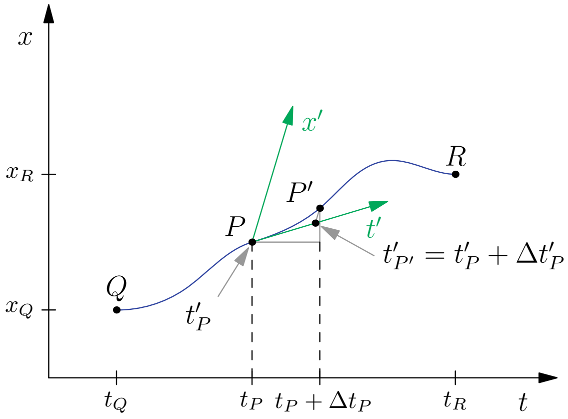


Fig. 9.23 Bob's trajectory and proper time

Fig. 9.23). The trajectory starts at event Q and ends at R. Obviously, the velocity at every moment always has to be smaller than the speed of light, i. e., $-c < v_B(t) < c$, where $v_B(t) = \mathrm{d}x_B(t)/\mathrm{d}t$ is the instantaneous velocity at time t.

How does the time pass for Bob? Suppose that, at Q, Alice's clock shows the time t_Q and Bob's clock shows t'_Q. Which time is shown by the clocks at event R?

Consider a small section of the trajectory that starts at event P and ends at event P' and that is sufficiently small to be able to consider Bob's velocity as constant, i. e., $v_B(t) = v_{B,P}$. For Alice, this interval lasts from t_P to $t_{P'}$, while, for Bob, it lasts from t'_P to $t'_{P'}$. Again for Alice, the interval has the length $\Delta t_P = t_{P'} - t_P$ (the index P indicates the fact that the interval starts at event P), while, for Bob, it has the length $\Delta t'_P = t'_{P'} - t'_P$. Then, we know that, in this time interval, as a consequence of time dilation, and from Alice's point of view, for Bob, the time

$$\Delta t'_P = \gamma^{-1}(v_B(t_P)) \cdot \Delta t_P = \sqrt{1 - v_B^2(t_P)/c^2} \cdot \Delta t_P \tag{9.11}$$

has passed.

The remainder is a simple integration exercise. For all of Alice's time intervals between t_Q and t_R, we must sum up the time that has passed for Bob (again, from Alice's point of view) in each of these time intervals. The sum corresponds to the time $t'_R - t'_Q$ that has passed for Bob on the trajectory from Q to R:

$$\begin{aligned} t'_R &= t'_Q + \int_{t_Q}^{t_R} dt' \\ &= t'_Q + \int_{t_Q}^{t_R} \sqrt{1 - v_B^2(t)/c^2}\, dt. \end{aligned}$$

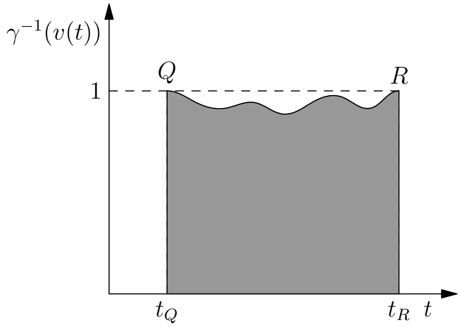


Fig. 9.24 Relativistic factor $\gamma^{-1}(v(t))$ of Bob (as seen by Alice) during his journey

Here, we have used $dt' = \sqrt{1 - v_B^2(t)/c^2}\, dt$, which is (9.11) for an infinitesimal small interval length Δt_P.

If we render the interval end t_R variable, the integral yields the time $t'(t)$ that Bob's clock shows when Alice's clock shows the time t, simultaneous for Alice (for this sake, we have replaced the integration parameter t with τ):

$$t'(t) = t'_Q + \int_{t_Q}^{t} \sqrt{1 - v_B^2(\tau)/c^2}\, d\tau. \tag{9.12}$$

To recognize the importance of the time t' associated with (9.12) to Bob's trajectory, let us come back to (9.11). The length $\Delta t'_P$ of the interval from P to P' is (from Alice's point of view) given by (9.11) for Bob. By squaring and multiplying with c^2, one gets

$$\left(c\Delta t'_P\right)^2 = \left(c^2 - v_B^2(t_P)\right)(\Delta t_P)^2 = (c\Delta t_P)^2 - (v_B(t_P)\Delta t_P)^2.$$

The expression $\Delta x := v_B(t_P)\Delta t$ corresponds exactly to the difference of the x-coordinates of P and P'. Therefore, we can write

$$\left(c\Delta t'_P\right)^2 = (c\Delta t_P)^2 - (\Delta x_P)^2 \equiv (\Delta s)^2.$$

From Sect. 9.5, we know that the right side of this equation is an invariant, namely, the spacetime distance of the events P and P'. For other inertial observers, therefore, the quantity $(\Delta s)^2$ has the same value as for Alice. Therefore, *the time period given by* (9.12) *is independent of Alice*. If another inertial observer were to carry out this procedure (the integral), he would assign the same times t' to the events on Bob's trajectory. The time t' associated in this way with Bob's trajectory is called Bob's **proper time**. It is the time that a clock carried by Bob shows (and therefore also Bob's age). Note that many authors use the Greek letter τ for the proper time and, e. g., write $\tau(t)$ in (9.12). The reason for doing so is that t' is not a coordinate of an inertial frame.

Now, we come back to the **twin paradox** and to Albert and Bert. Let Bert's trajectory start at Albert's location and also end there. Therefore, $x_B(t_Q) = x_B(t_R) = 0$. What is the time that passed for Bert on his journey? The answer follows directly from (9.12):

$$\Delta t' = t'_R - t'_Q = \int\limits_{t_Q}^{t_R} \sqrt{1 - v_B^2(t)/c^2}\, dt. \tag{9.13}$$

The integrand is shown in Fig. 9.24. The relation $\gamma^{-1}(v_B(t)) = \sqrt{1 - v_B^2(t)/c^2} \leq 1$ is valid in the whole time interval, whereas the equal sign is valid only if $v_B(t) = 0$. But this cannot hold in the whole interval, otherwise, Bert would not move away from Albert. The area under the curve in Fig. 9.24 therefore is smaller than $t_R - t_Q$, i. e.,

$$\Delta t' < t_R - t_Q \equiv \Delta t.$$

Independent of the form of his trajectory, for Bert, less time passes on his journey than for Albert. Bert ages more slowly than Albert.

Exercise 39: Calculate the proper time (9.13) for Bert if

$$v_B(t) = \begin{cases} +v & \text{if } t_Q \leq t < (t_Q + t_R)/2 \\ -v & \text{if } (t_Q + t_R)/2 \leq t \leq t_R \end{cases}$$

and discuss the result in view of the twin paradox.

9.11 Examples

9.11.1 Again: Muons

In the chapter on length contraction, we explained the experimental results of Frisch and Smith using the point of view of the muons and applied the length contraction to the trajectory traveled by them. How does an observer resting in the inertial frame attached to the Earth's surface explain the experiment there? There will be no length contraction, because the muon's trajectory rests relative to the experimenters. Indeed, the time dilation comes into play here. Suppose the muon carries a clock. Then, we know that the muon's time—from the observer's point of view—passes more slowly than for the muons themselves. The time dilation factor is given by $\gamma_v^{-1} < 1$. The average lifetime of the muon (for the observer) becomes larger than that of a resting muon. Therefore, we have to multiply the half-life t_M in (8.5) with γ and get exactly the same result as in (8.6).

We summarize: to explain the findings of Frisch and Smith, depending on the used inertial frame, we have to apply either length contraction or time dilation:

- **Point of view of the muon**
  - Flight distance: l/γ (because of length contraction)
  - Average lifetime of the muon: t_{M}
  - Factor in the e-function: $-l/(\gamma v t_{\mathrm{M}})$
- **Point of view of the observer resting on the Earth**
  - Flight distance: l
  - Average lifetime of the muon: γt_{M} (because of time dilation)
  - Factor in the e-function: $-l/(\gamma v t_{\mathrm{M}})$.

It is also possible to describe the experiment from the point of view of an inertial observer that moves relative to the Earth and the muon. However, this would be quite inconvenient, because both—length contraction and time dilation—have to be applied. But in the end, the result would be the same.

One sees again that the effects of special relativity are *spacetime effects* that, from one point of view, may be pure length contraction and, from another one, pure time dilation. In general, both effects will contribute, however.

9.11.2 The Experiment by Hafele and Keating

Introduction. To demonstrate time dilation, the physicist Joseph C. Hafele and the astronomer Richard E. Keating, in 1971, flew in an airliner around the world, once eastwards and once westwards, carrying a caesium atomic clock with them, a clock like that described in Sect. 9.4. After landing, the American scientists compared the clock's time with the time of another atomic clock of the same type that had been left back on Earth (Fig. 9.25).[31]

Hafele and Keating expected **two effects** to influence the operation of their clocks [HafeleKeating72a]. One is obvious: it is the **time dilation due to relative motion**, as discussed in Sect. 9.2. The time dilation, however, cannot be calculated directly from (9.2), because, in Hafele and Keating's experiment, the Earth's surface cannot be considered as an inertial frame. We will come back to this point in a minute. The other effect is **time dilation due to the gravitational field**, which also has an influence on the clocks. This effect, however, cannot be explained within the framework of special relativity, we have to borrow it from Einstein's **general theory of relativity** (GR).

Description of motion. In the calculations that we have made so far, the surface of the Earth was sufficiently close to an inertial frame. Now, this is not the case anymore.

[31] The scientists actually carried four atomic clocks with them and left four further ones back for comparison. The reason for this was to reduce systematic error by averaging them out.

Fig. 9.25 Hafele (right) and Keating (left) carrying out their experiment with their atomic clocks

But we can still consider the center of mass of the Earth as moving uniformly in an inertial frame. We choose the coordinate system such that its origin coincides with the Earth's center of mass and the coordinate axes point in directions that are fixed with respect to the starry sky (therefore, the coordinate system does not follow the rotation of the Earth). For our purpose, this reference frame is sufficiently close to an inertial frame. We will refer to it as *the inertial frame* and use the *Earth-centered inertial* (ECI) coordinate system. It has its origin in the mass center of the Earth and uses the equatorial plane as its reference plane. The ECI is the only inertial frame in this discussion, and we describe all effects from the point of view of this coordinate system.

For practical reasons, we introduce a further reference frame with the *Earth-centered Earth-fixed* (ECEF) coordinate system, which shares its origin and reference plane with the ECI, but which rotates with the Earth. In the ECEF, one uses the latitude, longitude and altitude to specify the location of a point. The ECEF is not an inertial frame.

We start by describing the motion of the observer and the planes in the inertial frame and suppose that everything happens on the Earth's equator. The observer is located on the equator and the planes fly along the equator. In the ECEF, the motion is simple and is shown in Fig. 9.26 on the left side. In the ECI, the planes and the observer all move in the eastwards direction, because the velocity of the observer due to the Earth's rotation is larger than that of the planes. The motion is shown in Fig. 9.26 on the right side. The red line in the figure shows the location where the planes land after flying around the Earth.

To make this quantitative, let φ_O be the angular position of the observer and ω_E the angular velocity of the Earth in the ECI. Furthermore, let $\omega_\mathrm{p} = v_\mathrm{p}/R_\mathrm{E}$ be the angular velocity of the plane relative to the observer, where v_p is the plane's velocity and R_E the radius of the Earth. Let us take $v_\mathrm{p} = 900\,\mathrm{km/h} = 0.25\,\mathrm{m/s}$.

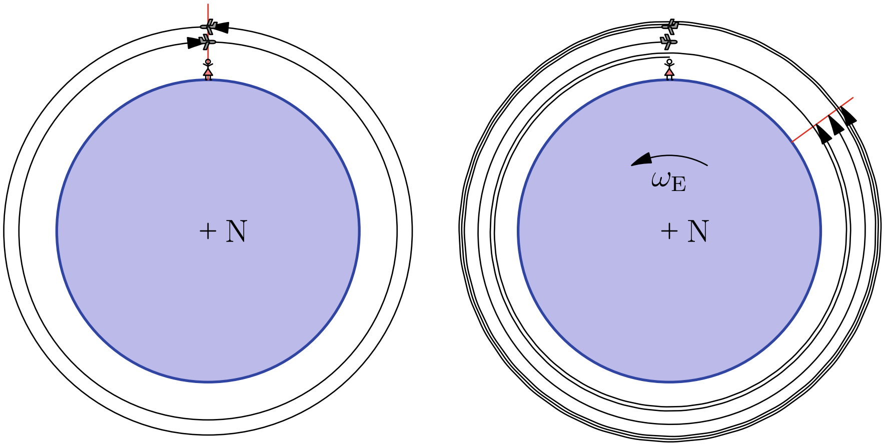


Fig. 9.26 Motion of the observer and the planes in the Earth-centered Earth-fixed coordinate system (left figure) and in the Earth-centered inertial coordinate system (right figure). We are looking along the Earth's rotation axis, from north to south, and the blue circle is the equator

In the ECI, the observer (or a clock at rest on the Earth's surface), due to the rotation of the Earth, has a velocity of about $v_O \approx 40{,}000\,\mathrm{km}/24\,\mathrm{h} = 0.46\,\mathrm{km/s}$.[32] We will denote this clock in the following as the *observer's clock*. Remember that the observer's reference frame is not an inertial frame.

Finally, let φ_+ and φ_- be the angular position of the eastwards and the westwards flying plane, respectively. Then, in the inertial frame, we have

$$\varphi_O(t) = \omega_E t \ ,$$
$$\varphi_\pm(t) = (\omega_E \pm \omega_p)t .$$

After having departed at $t = 0$ at the location $\varphi_O = 0$ of the observer, when do the planes arrive at the observer's location again? This is easy to see in the rotating frame (the ECEF coordinate system) where the observer is at rest. The *traveling time* t_0 is the same for the eastwards and the westwards flying planes and given by $t_0 = 2\pi R_E/v_p$. Plugging in numbers, we get $t_0 = 40{,}000\,\mathrm{km}/0.25\,\mathrm{km/s} = 160{,}000\,\mathrm{s} = 44.4\,\mathrm{h}$, and the Earth rotates almost twice while the planes are flying.

For the angular positions at $t = t_0$, in the inertial frame, we have

$$\varphi_\pm(t_0) = (\omega_E \pm \omega_p)\frac{2\pi R_E}{v_p} = \varphi_O(t_0) \pm 2\pi \ ,$$

which shows that the observer and the planes indeed meet at $t = t_0$.

[32] Actually, one rotation of the Earth takes a *sidereal day*, which is about 4 min less than a day. We take 24 h because this small difference is irrelevant to our calculation.

Proper time and time differences. At $t = 0$, the observer and the planes synchronize their clocks, which is easy because they are at the same location and we do not have to argue about the inertial frame in which the clocks are synchronous. Then the planes start their journey and accumulate proper time τ (see Sect. 9.10). At $t = t_0$, they meet again and compare their clock settings. We will consider the time differences

$$\tau_+ - \tau_\mathrm{O} = \begin{cases} \text{reading of clock of plane that traveled eastwards minus} \\ \text{reading of observer's clock} \end{cases},$$

$$\tau_- - \tau_\mathrm{O} = \begin{cases} \text{reading of clock of plane that traveled westwards minus} \\ \text{reading of observer's clock.} \end{cases}$$

The effects of relative motion and of gravitation can be considered independently.

Time dilation due to relative motion. First, we deal with the **effect of relative motion** of special relativity.

According to (9.13), we have

$$\begin{aligned} \tau_\mathrm{O} &= \gamma(v_\mathrm{O})^{-1} t_0 \\ \tau_\pm &= \gamma(v_\mathrm{O} \pm v_\mathrm{p})^{-1} t_0. \end{aligned}$$

All three clocks (that of the observer and those on the planes) run more slowly than the inertial clock. This is nothing but the twin paradox. The faster the clock, the more slowly it runs, therefore, we have $\tau_+ < \tau_\mathrm{O} < \tau_-$.

For the results of the clock comparisons, using the approximation $\gamma(v)^{-1} \approx 1 - v^2/2c^2$, which is valid for small velocities, we get

$$\begin{aligned} \tau_\pm - \tau_\mathrm{O} &= \left(\gamma(v_\mathrm{O} \pm v_\mathrm{p})^{-1} - \gamma(v_\mathrm{O})^{-1}\right) t_0 \\ &\approx \left[\left(1 - (v_\mathrm{O} \pm v_\mathrm{p})^2/2c^2\right) - \left(1 - v_\mathrm{O}^2/2c^2\right)\right] t_0 \\ &= \frac{1}{2c^2}\left(v_\mathrm{O}^2 - (v_\mathrm{O} \pm v_\mathrm{p})\right) t_0 \\ &= \frac{1}{2c^2}\left(\mp 2 v_\mathrm{O} v_\mathrm{p} - v_\mathrm{p}^2\right) t_0 \\ &= \mp\frac{1}{2c^2}\left(2v_\mathrm{O} \pm v_\mathrm{p}\right) v_\mathrm{p} t_0. \end{aligned}$$

Plugging in the numbers (velocities in km/s), we get

$$\tau_\pm - \tau_\mathrm{O} = \mp\frac{(2v_\mathrm{O} \pm v_\mathrm{p})v_\mathrm{p}}{2c^2} t_0 = \mp\frac{(2 \cdot 0.46 \pm 0.25) \cdot 0.25}{1.8 \times 10^{11}} \cdot 1.6 \times 10^5\ \mathrm{s}$$

or

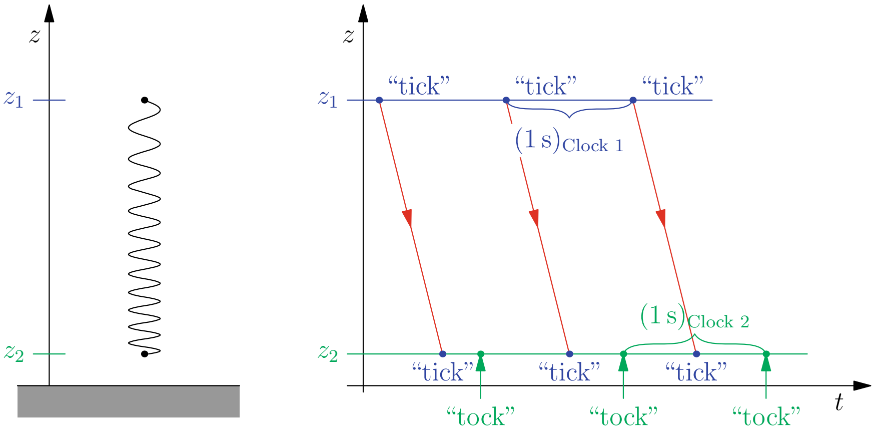


Fig. 9.27 Left: Frequency increase of a photon traveling down a gravitational field. Right: Clock period of one clock at height z_2 on the Earth (green, "tock") and another clock of the same type in a plane at height z_1 (blue, "tick"), as seen from the Earth. For the observer on Earth, 1 s as given by clock 1 takes less time than 1 s given by the local clock 2. Note that the plane here does not move relative to the observer on Earth

$$\tau_+ - \tau_O = -260\,\mathrm{ns}\,,$$
$$\tau_- - \tau_O = 149\,\mathrm{ns}.$$

For the observer, the clock on the eastwards-flying plane runs slow, whereas that on the westwards-flying plane runs fast. The latter fact may seem strange, as the clock in the eastwards-flying plane also moves relative to the observer. The reason for this is that the observer is not at rest in an inertial frame.

The values calculated for the actual flight trajectories are given in column "Prediction/Relative motion" in Table 9.1. The considerable deviations from our values are due to the fact that the planes did not travel on the equator.

Time dilation due to gravitation. Next, we deal with the time dilation caused by gravitation. Special relativity can only describe **gravitation** in an approximation that corresponds to the treatment of gravitation in classical mechanics. Actually, gravitation is a much more complicated physical phenomenon. The currently accepted theory of gravitation is **Einstein's general theory of relativity** (GR) from the year 1915. This theory has been spectacularly confirmed several times, for instance, with the motion of Mercury's perihelion, the bending of light rays that pass near massive objects (gravitational lenses), the time dilation due to the gravitational field (as discussed in the last section and this one), and, most impressively, by the detection of gravity waves.

According to general relativity, the "deeper" a clock is placed in a gravitational field, the more slowly the clock runs—even if the clocks are at rest relative to each other.[33] A clock on the Earth's surface runs more slowly than a clock on the Moon's surface, and a clock on the Sun's surface runs even more slowly than a clock on the Earth's surface. If, at the location P_1 of clock 1, the gravitational potential has the value U_1 and, at the location P_2 of clock 2, there's a gravitational potential U_2, then the difference in the clock rates is given by the formula

$$\frac{\Delta t_1}{\Delta t_2} = 1 + \frac{U_1 - U_2}{c^2}. \tag{9.14}$$

If clock 2 is "deeper" in the gravitational field than clock 1, then $U_1 - U_2 > 0$ and, therefore, according to the formula above, $\Delta t_1 > \Delta t_2$ which means that, while the time Δt_1 passes on clock 1, only the time $\Delta t_2 < \Delta t_1$ passes on clock 2. In other words: clock 2 runs more slowly than clock 1.

If the locations P_1 and P_2 are close to the Earth (this includes the flying plane) and P_1 is higher than P_2 by a height difference of $\Delta z = z_1 - z_2$, we have $U_1 - U_2 = g\Delta z$, with g being the gravitational acceleration at the Earth's surface, which is $g \approx 9.81\,\mathrm{m/s^2}$; for simplicity, we will use the value of $10\,\mathrm{m/s^2}$ in our calculations.

Note that the difference in clock speed here is not symmetric in the observer. The clock "deep" in the gravitational field sees the clock that is "less deep" in the gravitational field run faster while the latter sees the former run more slowly.

The derivation of (9.14) is performed within the framework of general relativity, but we can also motivate it without knowing this theory. The arguably easiest reasoning uses the concept of photons. As Einstein found out in 1905 when he investigated the photoelectric effect,[34] a monochromatic wave with a frequency ν in some way consists of light particles, later called photons, with an energy given by $E = h\nu$, where h is the Planck constant.

Now, for the gravitational time dilation, we suppose that such a photon travels from a location P_1 at height z_1 downward to the Earth's surface at height z_2, within the gravitational field (see Fig. 9.27 on the left side). Suppose further that, at P_1, it has the frequency ν_1, and therefore the energy $E_1 = h\nu_1$. Traveling down to location P_2 at height z_2, it gains the potential energy $mg\Delta z$ from the gravitational field. Here, the mass is the *relativistic inertial mass* of the photon and is given by Einstein's famous formula $E = mc^2$, which we will discuss in detail in Sect. 13.1. When the photon arrives at location P_2, it has the energy E_2 and frequency $\nu_2 = E_2/h$, given by

$$E_2 = E_1 + mg\Delta z = h\nu_1 + (h\nu_1/c^2)g\Delta z = h\nu_2.$$

Now, $g\Delta z$ is the difference of the gravitational potential, $g\Delta z = U_1 - U_2$, and therefore

$$\nu_2 = \nu_1 + \nu_1\frac{U_1 - U_2}{c^2} = \nu_1\left(1 + \frac{U_1 - U_2}{c^2}\right). \tag{9.15}$$

[33] The "deeper" one is within a gravitational field, the smaller the gravitational potential U is. The latter is comparable to the potential energy.

[34] This is the work for which he actually received the Nobel price, not the theory of relativity.

Table 9.1 The experiment by Hafele and Keating: predictions and measured values for the difference in the clocks' rates

Direction	Prediction			Measurement
	Relative motion	Gravitation	Total	
Eastwards (ns)	-184 ± 18	144 ± 14	-40 ± 23	-59 ± 10
Westwards (ns)	96 ± 10	179 ± 18	275 ± 21	273 ± 7

A positive value means that the clock transported in the airplane runs faster than the clock left at the airport. *Source* Loc. cited Science article by Hafele and Keating

Now, $\Delta t_1 = 1/\nu_1$ and $\Delta t_1 = 1/\nu_1$ are the clock periods and we recover (9.14).

For the actual case with our planes, we take $\Delta z = 10\,\text{km}$ and get

$$\begin{aligned}\tau_{\pm} - \tau_{\text{O}} &= \frac{U(z_1) - U(z_2)}{c^2}\tau_{\text{O}} \\ &\approx \frac{g\Delta z}{c^2}t_0 \\ &= \frac{10^5\,\text{m}^2/\text{s}^2}{9 \times 10^{18}\,\text{m}^2/\text{s}^2} \cdot 1.6 \times 10^5\,\text{s} = 1.78 \times 10^{-7}\,\text{s} \\ &= 178\,\text{ns}.\end{aligned}$$

The effect is the same for the westwards- and the eastwards-flying planes.

Thus, in our case, the clock in the airplane (which we imagine not moving relative to the Earth's surface) runs faster than that left behind at the airport. The values calculated for the actual flight trajectories[35] are given in the column "Prediction/Gravitation" in Table 9.1.

In the column "Predication/total", one finds the values for the *complete time dilation effect*, and in the column "Measurement", the experimentally measured values. Considering the error bar, the predicted values fit pretty well with the measured values (the measured values are from the original publication [HafeleKeating72b]). The experiment by Hafele and Keating is an impressive confirmation of the influences of relative motion in special relativity and of gravitation in general relativity to the "speed" of a clock.

Clock accuracy. The accuracy of atomic clocks has improved considerably over the years. In 1972, Hafele and Keating used Hewlett-Packard HP-5061A clocks, the successor model to HP's first commercial atomic clock. These clocks had an accuracy of about 10^{-11} (i.e., a maximum of 10^{-11} s of error in 1 s or 1 s of error in 10^{11} s $\approx$ 3179 years). Around the year 2000, the clocks used in the GPS satellites (also caesium beam clocks, like those used by Hafele and Keating) already had an accuracy, a 200-fold improvement, of about 5×10^{-14}. At this time, there already existed a much better design for caesium clocks, the *caesium fountain clocks* (which,

[35] The planes had different trajectories, and for this reason, the effect of gravitation is different for the two planes.

however, use free-falling atoms and do not work within a satellite in space). In the year 2007, the primary caesium clock of the American NIST, the NIST-F1, reached an accuracy of about 4×10^{-16}, 25,000 times better than the clocks of Hafele and Keating, and also likely representing the physical limits of this design. Soon after, laboratory versions of a completely new design, the *optical atomic clock*, became functional. These clocks are not based on microwave transitions in atoms (as in caesium), but on optical transitions. Around 2010, the NIST operated such a clock with an accuracy of 10^{-17}, one million times more accurate than Hafele and Keating's clock. With this clock, the Wineland group [Chou+10] at NIST was able to detect the time dilation with velocities as low as 10 m/s and height differences of just 1 m, confirming Einstein's theories. And the improvement of these clocks is still ongoing.

9.11.3 Satellite Navigation

An application that many people use each day is **satellite navigation**. Navigation systems in automobiles use satellite navigation, as do almost all smartphones. The most conversant satellite navigation system is the *Global Positioning System (GPS).*[36]

How it works. A satellite navigation system consists of a number of *satellites* in well-defined orbits around the Earth that send signals down to us. An electronic *receiver* on Earth receives signals from several of these satellites and is able to determine its own position from these signals. In the case of GPS, the receiver needs signals from at least four satellites to determine its position. GPS currently consists of almost 30 satellites that orbit the Earth twice a day, each satellite traveling along one of six different orbital planes. This configuration is chosen in such a way that, at each point on the Earth and at any moment, there are at least four satellites higher than 15° (called the elevation mask angle) over the horizon. In this way, it is ensured that the signals can be received with good quality, provided that the satellites are not blocked by buildings or the like. Actually, for most of the time, at least nine satellites are visible.

How does the determination of the position work with satellites? Let us first assume that the Earth-centered Earth-fixed (ECEF) system could be considered an inertial frame. This assumption is actually not appropriate, and we will correct this later.

Imagine a satellite S_1, which knows its exact position and the current time, in an orbit around the Earth. This satellite regularly[37] sends a data packet to Earth. This data packet contains the satellite's position $\boldsymbol{r}_{\mathrm{S},1}$ at the time $t_{\mathrm{S},1}$ when the data packet was sent, as well as this time itself. The receiver includes a clock that is synchronized with the clocks in the satellites, "synchronized" meaning relative to the considered inertial frame.

[36] The correct term is *Navigational Satellite Timing and Ranging—Global Positioning System* (NAVSTAR GPS).

[37] In GPS, this happens once every 30 s.

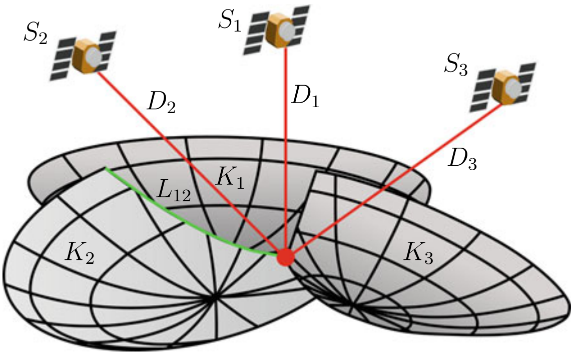


Fig. 9.28 Position determination with three satellites and synchronized clocks. Two spheres, in general, intersect at a circle (green) and three spheres, in general, intersect at two points (one of those is drawn in red)

When the receiver gets the data packet, it records the time of arrival t_R and can calculate the traveling time $t_R - t_{S,1}$ of the signal. Because of the principle of the absolute speed of light, from the traveling time, the distance $D_1 = (t_R - t_{S,1})/c$ of the satellite S_1 follows. Because the receiver is informed about the position $\boldsymbol{r}_{S,1}$ of the satellite, the receiver then knows that it is located on a sphere K_1 with radius D_1 around the satellite S_1 (see Fig. 9.28). Now imagine a second satellite S_2 of the same type. The receiver handles this one in the same way as the first satellite, and thus knows the position $\boldsymbol{r}_{S,2}$ of S_2 and the distance D_2 between the satellite and the receiver, and eventually a further sphere K_2 where the receiver resides. In total, the receiver then knows that it is located on the intersection of both spheres, which is the circle L_{12} (see Fig. 9.28). We take a third satellite S_3 and a sphere K_3, which intersects the circle L_{12} at two points. One of these points is usually on the Earth's surface or close to it (imagine a receiver in a plane) and the other point is very far from it. This allows the receiver to exclude one of the two intersection points. To conclude: by receiving the position and the sending time of the signal of three different satellites, the receiver with a synchronized clock can determine its position.

One of our assumptions is actually not true for satellite navigation systems. Receivers of satellite navigation systems do not have a clock that is sufficiently precise to be synchronized well with the satellites' clocks. To overcome this problem, one uses a further satellite. The solution then follows from a method that is basically the same as that above. But instead of the intersection of three spheres in three-dimensional space, the intersection of four "hyperspheres" in four-dimensional spacetime is used. Put differently, if $\boldsymbol{r}_{S,1}, \ldots, \boldsymbol{r}_{S,4}$ are the positions of the four satellites at the times $t_{S,1}, \ldots, t_{S,4}$ when the signals that arrive simultaneously at the receiver were sent, then we have four equations

$$c(t_{S,i} - t_R) = \left|\boldsymbol{r}_{S,i} - \boldsymbol{r}_R\right| \quad \text{where } i = 1, \ldots, 4. \tag{9.16}$$

From these, the position $\boldsymbol{r}_\mathrm{R}$ of the receiver and the time t_R of the arrival of the signals follows.[38]

If the receiver gets signals from more than four satellites, the additional data is used to improve the precision of the determined position of the receiver.

The role of relativity. Without the consideration of the effects of relativity, satellite navigation would not work.[39] First, to calculate the distance of the satellites using (9.16), the principle of the absolute speed of light (which is valid only in inertial frames) is explicitly used. The light travels from the satellite to the receiver with a velocity of $c = 299{,}792{,}458$ m/s, independent of the velocities of the satellite and the receiver.[40] Second, the influence of relative motion and gravitation on the clocks has to be taken into account as well, exactly as discussed in the Hafele-Keating experiment in Sect. 9.11.2. This is what we will do now.

Inertial frame and clock synchronization. But let us first correct the faulty assumption that the ECEF system was an inertial frame. At the end of Sect. 7.9.1, we saw that this assumption causes synchronization errors that, indeed, would render the GPS useless.

The good news is that we still can perform this synchronization using clocks resting on the Earth's surface (in the ECEF system). If we know the locations and velocities of a *master clock* on the Earth and of the satellite whose clock we want to set according to the master clock, we can send a pulse from the master clock to the satellite at t_0. The clock on the satellite then has to be set to $t_0 +$ (signal traveling time). The signal traveling time, however, will not be given by the distance between the master clock and the satellite, divided by the speed of light. The reason for this is that, in the accelerated ECEF system, the principle of the absolute speed of light does not hold and the speed of light will not be c. But this is not a problem, because we can perform the calculation of the signal traveling time (as "experienced" in the ECEF system) in an inertial frame, for instance, the Earth-centered inertial (ECI) frame discussed in Sect. 9.11.2 on the Hafele-Keating experiment. In the inertial frame, the principle of the absolute speed of light holds.

Influence of gravitation. In the discussion of the Hafele-Keating experiment, we learned that a clock "deeper" in the gravitational field runs more slowly than a clock that is not as "deep" in the field. We calculate this effect for the clocks on the satellites and that on the Earth.

The fraction of the clocks' speeds according to (9.14) amounts to $\Delta t_\mathrm{S}/\Delta t_\mathrm{E} = 1 + (U(R_\mathrm{S}) - U(R_\mathrm{E}))/c^2$. Here, Δt_S is the amount of time that passes on the satellite clock located at a distance R_S from the mass center of the Earth, while time Δt_E passes on the clock located at the Earth's surface, at a distance R_E from the mass center of the Earth.

[38] Here, we assume that the signals arrive at the same time. In reality, this is not the case, and therefore the receiver has to perform an additional mathematical correction on the result.

[39] The consequences of the relativistic effects for the GPS are very nicely discussed in [Ashby03].

[40] Indeed, the signal moves a little bit more slowly in the atmosphere, an effect that is taken into account in satellite navigation.

Now, our approximation $U_2 - U_1 = g\Delta z$ no longer works for a satellite, because the gravitational acceleration at the location of the satellite approximately 20,000 km above the surface of the Earth differs considerably from g. The gravitational potential U in a distance $r \geq R_{\mathrm{E}}$ from the center of mass of the Earth is given by $U(r) = -GM_{\mathrm{E}}/r$, where M_{E} is the Earth's mass and G Newton's gravitational constant.

To express GM_{E} with quantities accessible on the Earth's surface, we notice that the gravitational force acting on a test mass m at the Earth's surface is $F_{\mathrm{G}} = -mU'(R_{\mathrm{E}}) = GM_{\mathrm{E}}m/R_{\mathrm{E}}^2$. Because this must be equal to mg, we have $GM_{\mathrm{E}} = gR_{\mathrm{E}}^2$. Therefore, we can write the gravitational potential in the form $U(r) = -gR_{\mathrm{E}}^2/r$.

In total, the relative difference in the clocks' rates caused by gravitation is

$$\begin{aligned}\alpha = \left(\frac{\Delta t_{\mathrm{S}}}{\Delta t_{\mathrm{E}}} - 1\right) &= \frac{U(R_{\mathrm{S}}) - U(R_{\mathrm{E}})}{c^2} \\ &= \frac{gR_{\mathrm{E}}^2}{c^2}\left(\frac{1}{R_{\mathrm{E}}} - \frac{1}{R_{\mathrm{S}}}\right) = \frac{gR_{\mathrm{E}}}{c^2}\left(\frac{R_{\mathrm{S}} - R_{\mathrm{E}}}{R_{\mathrm{S}}}\right) \\ &= \frac{10\,\mathrm{m/s}\cdot 6{,}378\,\mathrm{km}}{(3\times 10^5\,\mathrm{km/s})^2}\cdot\frac{20{,}000\,\mathrm{km}}{26{,}378\,\mathrm{km}} \\ &\approx 5\times 10^{-10}.\end{aligned}$$

A more detailed inspection yields the value $\alpha = 4.4647 \times 10^{-10}$. This difference in the clocks' rates has to be corrected, otherwise, the clocks would already be off by 4×10^{-10} s after only one second. This seems small, but in this time, the light travels 12 cm. Just one second after synchronizing the clocks, the error in the position determination with satellites would already be on this order!

So, the goal is to keep the clocks on the satellite and those on Earth running at the same "speed", and one achieves this by using satellite clocks that are a bit miscalibrated. The atomic clocks left on the Earth tick with a frequency of $\nu_E = 10.23\,\mathrm{MHz}$ (this is not the frequency of the atomic transition used in the clock, but a derived frequency), while the satellite clocks *on the Earth* must tick with a frequency of $\nu_S = (1 - 4.4647 \times 10^{-10}) \cdot 10.23\,\mathrm{MHz} = 10.229\,999\,995\,43\,\mathrm{MHz}$. On the Earth, these modified clocks run slow but once in the satellite's orbit and as seen from the Earth, they run perfectly in sync with the clocks left on the Earth.

Exercise 40: Suppose the frequency of the satellite clocks was wrong by one decimal in the last digit of the frequency ν_S, i. e., $\nu_S = \ldots 995\,44$ instead of $\ldots 995\,43$. How much time would it take until the accumulated error corresponds to an error in distance of 1 m?

Chapter 10
Lorentzian Addition of Velocities

10.1 Introduction

In Sect. 3.5, we have shown how velocities are added in classical mechanics. By adding velocities, we mean the following: there are three inertial observers Alice, Bob and Claire. Bob moves with velocity v_{BA} relative to Alice and Claire moves with velocity v_{CB} relative to Bob (both move in the same direction). Then, we want to know the velocity v_{CA} of Claire relative to Alice. In classical physics, this addition is just a plain algebraic addition. In special relativity, this can no longer be correct, because, for example, for $v_{\mathrm{BA}} = v_{\mathrm{CB}} = 2c/3$, we would get $v_{\mathrm{CA}} = 4c/3$, which is larger than the speed of light, and therefore impossible. The basic ingredient for the derivation of the Galilean addition of velocities was the Galilei transformation, which gives us an indication of what we have to do differently.

In this chapter, we will use a graphical method to derive the relativistic addition of velocities. A nice application is the Fizeau experiment, where the speed of light in moving liquids is measured.

10.2 Addition of Velocities

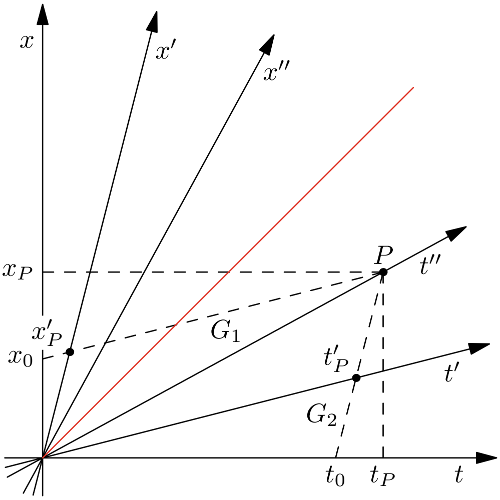
Let us now fulfill the promise given in Sect. 6.1 and derive how velocities are added in special relativity. To differentiate from the Galilean addition of velocities, we will talk about the **Lorentzian addition of velocities (LAV)**.[1]

For the derivation, we imagine three inertial observers Alice, Bob and Claire. In Alice's coordinate system, Bob moves according to $x_{\mathrm{B}}(t) = v_{\mathrm{BA}}t$ and Claire accord-

[1] In the literature, the notion the **velocity-addition formula** is usually used.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_10

Fig. 10.1 For the derivation of the Lorentzian addition of velocities



ing to $x_C(t) = v_{CA}t$ (see Fig. 10.1). Bob uses the primed and Claire the double-primed coordinate variables.

What we need now is the velocity v_{CB} of Claire in Bob's coordinate system. For small velocities, we simply have $v_{CB} = v_{CA} - v_{BA}$, but we will also allow for large velocities here.

Take an event P on Claire's time axis. The coordinates of this event, for Alice, are $P = (t_P, x_P)$. As mentioned, from Alice's point of view, Claire has the velocity

$$v_{CA} = \frac{x_P}{t_P}. \tag{10.1}$$

For Bob, the coordinates of P are given by $P = (t'_P, x'_P)$. Bob determines Claire's velocity to be $v_{CB} = x'_P/t'_P$. As shown in Fig. 10.1, through P, we draw Bob's lines of "equal position" G_1 and simultaneity G_2. These lines intersect Alice's coordinate axes at x_0 and t_0. Because of $t'_P = \gamma_{BA}t_0$ and $x'_P = \gamma_{BA}x_0$ (where $\gamma_{BA} := \gamma(v_{BA})$), we also have

$$v_{CB} = \frac{x'_P}{t'_P} = \frac{x_0}{t_0}. \tag{10.2}$$

Now, we express the coordinates x_P, t_P as functions of x_0, t_0. Then, using (10.1) and (10.2), we can construct a relation between the relative velocities v_{CA} and v_{CB}.

The equations of the lines G_1 and G_2 are

$$G_1 : x = x_0 + v_{BA}t \ ,$$
$$G_2 : x = \frac{c^2}{v_{BA}}(t - t_0).$$

They intersect at the event (t_P, x_P), for which the following holds:

$$x_P - v_{\mathrm{BA}} t_P = x_0 \ ,$$
$$\frac{v_{\mathrm{BA}}}{c^2} x_P - t_P = -t_0 .$$

Solving for x_P and t_P yields, for this event,

$$x_P = \alpha \cdot (x_0 + v_{\mathrm{BA}} t_0) \ ,$$
$$t_P = \alpha \cdot \left(\frac{v_{\mathrm{BA}}}{c^2} x_0 + t_0 \right) \ ,$$

with $\alpha = c^2/(v_{\mathrm{BA}}^2 - c^2)$. Because of (10.1) and (10.2), this results in

$$v_{\mathrm{CA}} = \frac{x_P}{t_P} = c^2 \frac{v_{\mathrm{CB}} + v_{\mathrm{BA}}}{c^2 + v_{\mathrm{CB}} v_{\mathrm{BA}}} = \frac{v_{\mathrm{CB}} + v_{\mathrm{BA}}}{1 + \dfrac{v_{\mathrm{CB}} v_{\mathrm{BA}}}{c^2}} .$$

Lorentzian addition of velocities: If Bob moves with velocity v_{BA} relative to Alice and Claire with velocity v_{CB} relative to Bob, then Claire moves relative to Alice with the velocity v_{CA}, given by

$$v_{\mathrm{CA}} = v_{\mathrm{CB}} \oplus v_{\mathrm{BA}} := \frac{v_{\mathrm{CB}} + v_{\mathrm{BA}}}{1 + \dfrac{v_{\mathrm{CB}} v_{\mathrm{BA}}}{c^2}} . \tag{10.3}$$

One immediately recognizes the limit of classical mechanics: if $v_{\mathrm{CB}} \ll c$ and $v_{\mathrm{AB}} \ll c$, then the denominator is approximately equal to one and one gets the Galilean addition of velocities. (Here, only for velocities that point in the same direction. The general relativistic case is more complicated.)

If one of the velocities, e.g., v_{CB}, is equal to the speed of light, we get

$$v_{\mathrm{CA}} = c \oplus v_{\mathrm{BA}} = \frac{c + v_{\mathrm{BA}}}{1 + \dfrac{c v_{\mathrm{BA}}}{c^2}} = c .$$

Altogether, we have

$$v_{\mathrm{CA}} = v_{\mathrm{CB}} \oplus v_{\mathrm{BA}} \approx v_{\mathrm{CB}} + v_{\mathrm{BA}} \quad \text{if } v_{\mathrm{CB}}, v_{\mathrm{BA}} \ll c,$$
$$v_{\mathrm{CA}} = c \oplus v_{\mathrm{BA}} = c \ ,$$
$$v_{\mathrm{CA}} = c \oplus c = c .$$

Exercise 41: Show that, resolving (10.3) for v_{CB}, corresponds to exchanging v_{CB} for v_{CA}, and making the replacement $v_{\mathrm{BA}} \to -v_{\mathrm{BA}}$. This is the same as exchanging Alice for Bob.

Exercise 42: Show that, for $x, y \in I$ and $I = [0, 1]$, the quantity $z = (x + y)/(1 + xy)$ also lies in the interval I. The easiest way to do this is to calculate $z - 1$, expand it to one fraction and recognize a complete square in the nominator. Show in this way that, for $0 \leq v_{\mathrm{CB}}, v_{\mathrm{BA}} \leq c$, the relation $0 \leq v_{\mathrm{CA}} \leq c$ holds. By adding two velocities that are not larger than c, one cannot get a velocity larger than c.

Exercise 43: The relativistic formula for adding velocities defines a binary operation $(a, b) \mapsto a \oplus b$ with $a \oplus b := (a + b)/(1 + ab)$. Show that this operation has the following properties:

- Commutativity[2]: $a \oplus b = b \oplus a$
- Associativity: $(a \oplus b) \oplus c = a \oplus (b \oplus c)$
- Unique existence of an inverse element: $a \oplus b = 0$ implies $b = -a$.

Exercise 44: To describe relative motion, instead of the velocity, one can also use the **rapidity** $\theta = \operatorname{atanh}(v/c)$. For small velocities, $\theta \approx v/c$, while for large velocities $v \to c$, the rapidity goes to infinity.

Show that, while one has to use (10.3) to add velocities, the rapidities can just be algebraically added. If Bob moves with rapidity θ_{BA} relative to Alice and Claire with rapidity θ_{CB} relative to Bob, then Claire moves relative to Alice with rapidity

$$\theta_{\mathrm{CA}} = \theta_{\mathrm{CB}} + \theta_{\mathrm{BA}}.$$

Exercise 45: Show that, even with signals that do not move with the speed of light, one can synchronize clocks.

If one uses the classical formula for the addition of velocities, one finds a result that is consistent with the absolute simultaneity of Newton. But if one uses the relativistic formula for the addition of velocities, one arrives at the relativity of simultaneity, exactly like when we were synchronizing clocks with light pulses.

10.3 Digression: The Fizeau Experiment

The experiment. In optics, we are told that light in a medium moves more slowly than in vacuum. This, in the end, is the reason for the diffraction of light at interfaces.

[2] The Lorentz transformation is only commutative if the velocities that are added are parallel.

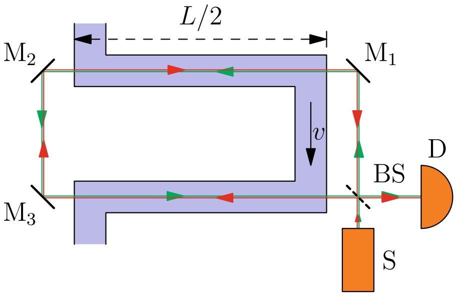


Fig. 10.2 Design of the Fizeau experiment. For an explanation, see the text

In an important class of media,[3] the speed of light c', as in vacuum, is independent of position and direction. In such media, the speed of light is given by $c' = c/n$. The *index of refraction n* usually depends on the frequency ω of the light, which is why you see colors if you send white light into a prism and why there is such a thing as a rainbow. The relation $c' = c/n$ was verified by the Frenchman Hippolyte Fizeau in 1849 for light in water. For light in water, n lies between 1.32 (for red light) and 1.35 (for blue light), corresponding to a light velocity of about 227,000 km/s (red) and about 222,000 km/s (blue).

But we now deal with another of the Frenchman's experiment, the **Fizeau experiment**, which he used in 1851 to measure the speed of light in **moving water**. His experiment is similar to that that Michelson and Morley carried out 30 years later. In both, light from the same source travels different paths and is then brought to interference. From the change of the interference pattern, the change of the phase difference of the interfering waves can be deduced. The change of the phase difference then gives us length differences in fractions of the wavelength of the used light (these are length differences on the order of 100 nm, which is 10,000 times smaller than a millimeter!), or, eventually, velocity differences.

The design of the experiment is shown in Fig. 10.2. A liquid medium (here, water) is flowing in glass tubes and has the flow velocity v. Light, by means of mirrors, is guided in different directions through this moving medium.

The monochromatic light wave of frequency ν emitted by the light source S falls onto a semitransparent mirror BS (which acts as a beam splitter), where it is split into two partial waves of equal intensity.

One partial wave is transmitted towards the semitransparent mirror, first traveling from mirror M$_1$ to mirror M$_2$ against the direction of flow through the medium, and then between mirror M$_3$ and the semitransparent mirror afterwards. It is then transmitted at the semitransparent mirror and travels to the detector D.

The other partial beam, after leaving the source, is reflected at the semitransparent mirror BS and moves on the way to mirror M$_3$ in the flow direction through the moving medium. On the way from mirror M$_2$ to M$_1$, it does the same again. Having arrived

[3] Such media in optics often are called *homogeneous media*. In theoretical physics, they would be called *homogeneous and isotropic*, because their properties are *independent of position and direction*. You see: even in physics, not everything is consistent : -) .

at the semitransparent mirror, the partial beam is reflected and interferes with the other partial beam on the way to the detector.[4] If the moving medium actually is at rest, one gets a certain interference pattern, and this pattern changes if one alters the flow velocity. Via the change of the interference pattern (actually the intensity at the detector), the phase difference of the partial beams is measured.

Explanation with special relativity. Both beams pass along a path of length L through the moving medium. The number of wavelengths that fit into this distance is L/λ' and the phase change between entry and exit is $2\pi L/\lambda'$. The wavelength λ' *in the medium* is related to the (fixed) light frequency ν via the light velocity c' in the medium through $\lambda'\nu = c'$. The phase change accumulated while traveling through the moving medium is related to the light velocity c' by $2\pi\nu L/c'$.

Let c'_+ be the velocity of the light in the moving medium in the direction of flow and c'_- that against the direction of flow. Further, let $\Delta\varphi_+ = 2\pi\nu L/c'_+$ and $\Delta\varphi_- = 2\pi\nu L/c'_-$ be the respective phase changes (which, via the light velocity, depends on the water's flow velocity). The phase difference $\Delta\varphi$ relevant for the interference depends on the light velocity in the medium via

$$\Delta\varphi = \Delta\varphi_- - \Delta\varphi_+ = 2\pi\nu L\left(\frac{1}{c'_-} - \frac{1}{c'_+}\right).$$

If the medium is at rest, we have $c'_+ = c'_- = c'$, and therefore $\Delta\varphi = 0$.

What do we expect for the light velocity in the moving medium? For an observer moving with the medium, the light propagates with the velocity $c' = c/n$. This observer moves relative to the experiment with the velocity $v \ll c'$. Therefore, we only have to add the velocities. Classically, we have $c'_+ = c' + v$ and $c'_- = c' - v$. Relativistically, we use the Lorentzian addition of velocities and get

$$\begin{aligned} c'_+ &= \frac{c'+v}{1+c'v/c^2} \approx c'\left(1+\frac{v}{c'}\right)\left(1-\frac{c'v}{c^2}\right) \\ &= c'\left(1+\frac{v}{c'}-\frac{c'v}{c^2}-\frac{v^2}{c^2}\right) = c'\left(1+\left(1-\frac{c'^2}{c^2}\right)\frac{v}{c'}-\frac{v^2}{c^2}\right). \end{aligned}$$

The second term in the parenthesis is of order v/c and the third is proportional to $(v/c)^2$. We neglect the third one and get

$$c'_+ \approx c' + \left(1-\frac{c'^2}{c^2}\right)v = c' + \left(1-\frac{1}{n^2}\right)v\,,$$

instead of the classical result $c'_+ = c' + v$. If the light travels against the direction of flow of the water, one gets, in the same way,

[4] Note that, at the detector D, only half of the intensity of the light wave arrives. When the partial beams come back to the semitransparent mirror BS, each of them is again split into two partial beams. Two of the resulting four partial beams travel to the detector and the remaining two go back to the light source. See also Footnote 1 in Sect. 5.2.1

$$c'_- = \frac{c' - v}{1 - c'v/c^2} \approx c' - \left(1 - \frac{1}{n^2}\right) v.$$

The results can be summarized by

$$c'_\pm = c' \pm \alpha_n v \ ,$$

where $\alpha_n = 1 - 1/n^2$ is **Fresnel's drag coefficient** (which is a historical designation, see also (5.7)). In the case of the Fizeau experiment, when one adds a large velocity $c/n < c$ and a small one $v \ll c$, the Lorentzian addition of velocities looks similar to the Galilean addition of velocities if one takes the flow velocity v of the medium only with a factor of $\alpha_n < 1$ into account. Note that, whereas special relativity yields $\alpha_n = 1 + 1/n^2$, classical physics gives us $\alpha_n = 1$.

The expected phase difference then becomes

$$\begin{aligned}
\Delta\varphi &= 2\pi \nu L \left(\frac{1}{c'_-} - \frac{1}{c'_+}\right) \\
&= \frac{2\pi \nu L}{c'} \left(\frac{1}{1 - \alpha_n v/c'} - \frac{1}{1 + \alpha_n v/c'}\right) \\
&\approx \frac{2\pi \nu L}{c'} \left(\left(1 + \alpha_n v/c'\right) - \left(1 - \alpha_n v/c'\right)\right) \\
&= 4\pi \nu L \alpha_n \frac{v}{c'^2}.
\end{aligned}$$

Using the wavelength $\lambda_0 = c/\nu$ in vacuum, we can write

$$\Delta\varphi = 4\pi \frac{L}{\lambda_0} \frac{v}{c} n^2 \alpha_n = 4\pi \frac{L}{\lambda_0} \frac{v}{c} \cdot \begin{cases} n^2 & \text{in the classical case} \\ n^2 - 1 & \text{in the relativistic case.} \end{cases} \tag{10.4}$$

We recognize the following: the larger the path length L and the higher the velocity of flow v of the medium, the larger the phase difference of the two partial beams.

Fizeau's finding. Fizeau, in his experiment, confirmed the Formula (10.4). His result is consistent with the relativistic prediction and in contradiction to the classical one. He has shown that the "naive" Galilean addition of velocities is wrong for the calculation of the velocity of light in moving media. His result confirms special relativity.

Exercise 46: We carry out the Fizeau experiment with the light of the (vacuum) wavelight $\lambda_0 = 589$ nm (the yellow of a sodium lamp). Then, the index of refraction of water is $n = 1.33$. And if $\Delta\varphi = 2\pi/100$ is the detection limit and $L = 1$ m, the water must move at least with the velocity $v \approx 1.15$ m/s, otherwise, the effect is too small for detection. Show this.

Chapter 11
The Lorentz Transformation: Derivation

In Sect. 3.4.2, we discussed the Galilei transformation and determined that it cannot be valid anymore for large velocities. The necessary consequence is that classical mechanics must be replaced with Einstein's relativistic mechanics (i. e., the special theory of relativity) and the Galilei transformation with the **Lorentz transformation** (LT) . In the same way as the Galilei transformation leaves the equations of classical mechanics form-invariant,[1] the Lorentz transformation leaves the equations of relativistic mechanics form-invariant and, in addition, does the same for the equations of electrodynamics.

This chapter is exclusively dedicated to the derivation of the Lorentz transformation. In Sect. 11.1, we derive the Lorentz transformation with our geometrical methods. For an even deeper understanding, in Sect. 11.3, we present an alternative and purely algebraic derivation.

In the next chapter, we will demonstrate the power of the Lorentz transformation and, using our new tool over a few lines, derive the effects of the relativity of simultaneity, time dilation, and length contraction.

11.1 Graphical Derivation of the Lorentz Transformation

The Lorentz transformation is the link between the coordinates of an arbitrary event E in both Alice's and Bob's coordinate systems. It includes all of the relativistic effects that we have discussed so far (as we will see in Sect. 12.1).

With the experience that we gained with geometric constructions in the meantime, it is easy to derive the Lorentz transformations. We consider two inertial observers Alice and Bob with their coordinate systems in standard configuration and an event E

[1] Form-invariant means that the equations keep their form when transformed.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_11

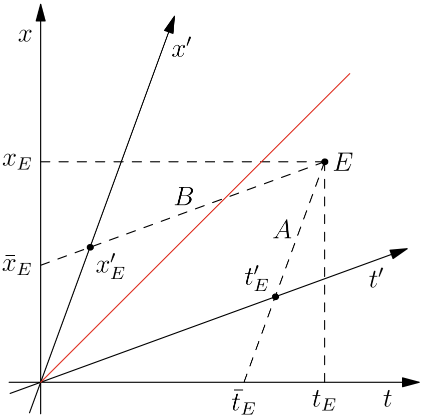


Fig. 11.1 Geometric derivation of the Lorentz transformation

(see Fig. 11.1). From Alice's point of view, this event has the coordinates (t_E, x_E), and for Bob, (t'_E, x'_E). We want to calculate Bob's coordinates of the event E from Alice's coordinates of it. This amounts to finding the functions f and g in $x'_E = f(t_E, x_E)$ and $t'_E = g(t_E, x_E)$.

To achieve this, we draw a line B through event E parallel to Bob's t'-axis. This line intersects Alice's x-axis at $\bar{x}_E$. In the same way, the line A through E and parallel to Bob's x'-axis intersects Alice's t-axis in $\bar{t}_E$ (note that $\bar{t}_E, \bar{x}_E$ are not coordinates of E!). We know already from Sect. 9.5 (see e. g., Fig. 9.14) that the barred variables are related to Bob's coordinates of E via $t'_E = \gamma_v \bar{t}_E$ and $x'_E = \gamma_v \bar{x}_E$. Now, we must find out how $\bar{t}_E$ and $\bar{x}_E$ depend on Alice's coordinates of E.

The line A is given by $(x - x_E) = (c^2/v)(t - t_E)$. With $x = 0$, we get $\bar{t}_E = t_E - (v/c^2)x_E$. The line B is given by $(x - x_E) = v(t - t_E)$. With $t = 0$, we get $\bar{x}_E = x_E - vt_E$. Now, we use the relations between the barred quantities to t'_E and x'_E discussed in the last paragraph and get the **Lorentz transformation**

$$\begin{aligned} t' &= \gamma_v \left(t - \frac{v}{c^2}x\right), \\ x' &= \gamma_v \left(x - vt\right). \end{aligned} \tag{11.1}$$

Here, we dropped the index E because the event is arbitrary.

The Lorentz transformation is a generalization of the Galilei transformation (3) where the former also works for large velocities. Therefore, it must be equal (or close to equal) for small velocities, where the Galilei transformation (3) is correct. This limiting case is easy to see. Using $v/c \to 0$, we get $\gamma_v = 1$ and $v/c^2 = 0$, and the Lorentz transformation (11.1) indeed becomes equal to the Galilei transformation (3).

Exercise 47: Show, using the Lorentz transformation (11.1), that $c^2t'^2 - x'^2 = c^2t^2 - x^2$ holds. Therefore, the hyperbola is independent of the coordinate system. It is an invariant of the Lorentz transformation.

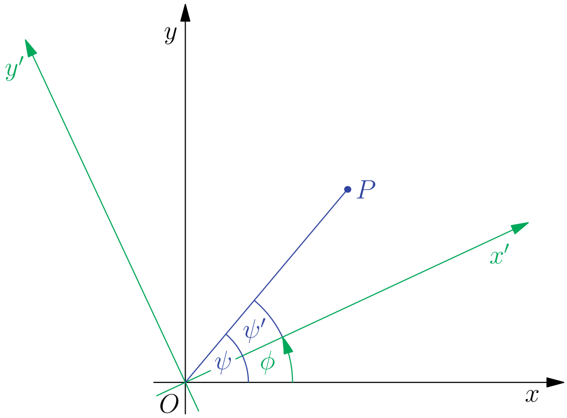


Fig. 11.2 To the derivation of the rotation in Euclidean space (see text)

11.2 Digression: The Lorentz Transformation in Matrix Form

Rotations in Euclidean space. We start by considering **rotations in Euclidean space**.

A rotation in a plane always leaves exactly one point fixed. We identify this point with the origin of our (orthonormal) coordinate system K with the coordinates x and y. In addition to this, we consider a further coordinate system K' with the coordinates x' and y' that shares its origin with K. Let K' be rotated in the positive mathematical direction (counterclockwise) by the angle ϕ relative to K (see Fig. 11.2).

Our task now is to calculate the coordinates (x', y') of an arbitrary point P, given its coordinates (x, y) in coordinate system K. If the line segment $\overline{OP}$ has the length r and encloses an angle ψ with the x-axis, we can give the unprimed coordinates immediately: $x = r\cos\psi$ and $y = r\sin\psi$. The primed coordinates follow in the same way: let ψ' be the angle enclosed by the line segment $\overline{OP}$ and the x'-axis, then, $x' = r\cos\psi'$ and $y' = r\sin\psi'$. Now, the relation between the angles is $\psi' = \psi - \phi$, and from the angle sum and difference identities of trigonometry, it follows that

$$\begin{aligned}\cos\psi' &= \cos(\psi-\phi) = \cos\psi\cdot\cos\phi + \sin\psi\cdot\sin\phi\ ,\\ \sin\psi' &= \sin(\psi-\phi) = \sin\psi\cdot\cos\phi - \cos\psi\cdot\sin\phi\ ,\end{aligned}$$

the left sides of these equations being exactly x'/r and y'/r. We can multiply the equations with r and, on the right sides, put the expressions for x and y. This implies that

$$\begin{aligned}x' &= x\cos\phi + y\sin\phi\ ,\\ y' &= -x\sin\phi + y\cos\phi\end{aligned}\qquad(11.2)$$

or, written in matrix notation,

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

This is how the coordinates (x, y) of an arbitrary point P transform when the coordinate system is rotated by an angle ϕ.

All pairs of physical quantities (p, q) that transform like x and y in (11.2) (or the generalization to all three space dimensions) are called **vectors** .

The distance of point P from the origin obviously has to be the same in both coordinate systems. Therefore,

$$r'^2 = x'^2 + y'^2 = x^2 + y^2 = r^2 . \tag{11.3}$$

For this particular reason, the quantity $r^2 = x^2 + y^2$ is called an *invariant of the rotation*. Invariant because, in a rotation, its value is unchanged.

Exercise 48: Prove (11.3) by substituting x', y' from the expression for the rotation (11.2) into the expression $x'^2 + y'^2$.

Exercise 49: Which geometric figure does not change in a rotation? One says that is in invariant under rotation.

The Lorentz transformation. Now, it is easy to see that the Lorentz transformation (11.1), in matrix notation, looks like

$$\begin{pmatrix} t' \\ x' \end{pmatrix} = \gamma_v \begin{pmatrix} 1 & -v/c^2 \\ -v & 1 \end{pmatrix} \begin{pmatrix} t \\ x \end{pmatrix} .$$

The matrix is actually symmetric when we "reflect" it at the diagonal line built by the two matrix entries "1". This can be seen if, instead of the time t, we use the quantity ct, which has the same unit as x. Then,

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \gamma_v \begin{pmatrix} 1 & -v/c \\ -v/c & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix} .$$

The invariant of the Lorentz transformation that corresponds to the radius in Euclidean rotations (see (11.3)) is already known to us from Sect. 9.5. It is the spacetime distance (9.2).

All pairs of physical quantities (p, q) that transform like ct and x in (11.1) (or the generalization to all four spacetime dimensions) are called **four-vectors** (or **4-vectors**). And for each of these four-vectors (p, q), the quantity $p^2 - q^2$ is **invariant** in a Lorentz transformation.

11.3 Digression: Analytic Derivation of the Lorentz Transformation

We now give a purely analytical derivation of the Lorentz transformation. It shows very nicely the inputs that lead to the Lorentz transformation. Be warned that the derivation is a bit lengthy, but, on the other hand, the logic is clear and the calculations are simple.

In contrast to the geometric derivation, we consider all four spacetime dimensions now. As usual, we use the unprimed variables (t, x, y, z) for Alice's coordinate system and the primed ones (t', x', y', z') for Bob's coordinate system. The Lorentz transformation determines the coordinates (t', x', y', z') of an event E in Bob's coordinate system when its coordinates (t, x, y, z) in Alice's coordinate system are given.

We choose the coordinates in standard configuration, i. e., the x- and x'-axis coincide and the y- and y'-axis are mutually parallel, as are the z- and the z'-axis. Furthermore, we shift the axes such that, at $t = t' = 0$, both spatial coordinate systems coincide.

First of all, Bob's coordinates could be arbitrary functions of Alice's coordinates, e. g., $x' = f(x, y, z, t)$, $y' = g(x, y, z, t)$, etc. These functions, however, can be quickly and considerably simplified.

The important argument here is that, according to Newton's law of inertia, particles upon which no force acts move with constant velocity vector in inertial frames, and thus its world line is a (straight) line. The Lorentz transformation, therefore, has to map lines to lines,[2] and thus, it has to be linear, i. e.,

$$\begin{aligned}
t' &= L_{00}t + L_{01}x + L_{02}y + L_{03}z \ , \\
x' &= L_{10}t + L_{11}x + L_{12}y + L_{13}z \ , \\
y' &= L_{20}t + L_{21}x + L_{22}y + L_{23}z \ , \\
z' &= L_{30}t + L_{31}x + L_{32}y + L_{33}z \ .
\end{aligned}$$

Thus, with this argument, we have simplified the four arbitrary functions to four linear functions. The coefficients L_{ij} (with $i, j = 0, 1, 2, 3$) may still depend on the velocity, but not on the coordinates.

Incorporating the choice of coordinates. The choice of the coordinates (standard configuration) simplifies the transformation considerably. First, $y = 0$ must imply $y' = 0$, for arbitrary values of t, x, z. This means that $L_{20} = L_{21} = L_{23} = 0$, and therefore $y' = L_{22}y$. The analogous observation holds for the z-coordinate. Next, the plane $x = vt$ must map to $x' = 0$ for arbitrary values of y, z. Therefore, y and z must not appear in the equations for t' and x', i. e., $L_{02} = L_{03} = L_{12} = L_{13} = 0$.

Furthermore, Bob moves with velocity v relative to Alice, i. e., $x = vt$ must imply $x' = 0$, and vice versa. For all events with $x = vt$, we get, from the equations

[2] Actually, it only has to be affine, and it can include a shift in addition to the linear transformation. This, however, is irrelevant, because we assume that the origin of the coordinate system can be chosen arbitrarily (spacetime is homogeneous).

above, $x' = L_{10}t + L_{11}vt = (L_{10} + vL_{11})t \overset{!}{=} 0$. This is the case if $L_{10} = -vL_{11}$, and therefore $x' = L_{11}(x - vt)$.

In this way, we arrive at

$$\begin{aligned} t' &= L_{00}t + L_{01}x \ , \\ x' &= L_{11}(x - vt) \ , \\ y' &= L_{22}y \ , \\ z' &= L_{33}z \ . \end{aligned}$$

In addition to what we have discovered so far, we also know that the constants L_{00}, L_{11}, L_{22}, and L_{33} all must be larger than zero. The reason for this is the relative orientation of the coordinate axes: the x'-axis points in the same direction as the x-axis and not in the opposite direction. And the same holds for the other three pairs of axes.

Now, we have to determine the remaining five coefficients.

Transversal direction. As a next step, we focus on how the *directions transversal (perpendicular) to the relative velocity* transform. For this purpose, let us consider a transformation, called an *xz-reversal*, which does the following:

$$x \leftrightarrow -x' \ , \quad y \leftrightarrow y' \ , \quad z \leftrightarrow -z' \ , \quad t \leftrightarrow t' \ .$$

This amounts to exchanging Alice and Bob and rotating their coordinate systems by an angle of 180° around the y- or y'-axis, respectively. Then, according to the relativity principle, the Lorentz transformation must be the same. The xz-reversal applied to $y = L_{22}y'$ produces $y' = L_{22}y$ (note that the dependency of L_{22} on the velocity is not a problem, as the relative velocity is the same after the xz-reversal), and therefore $L_{22}^2 = 1$. For $v \to 0$, the transformation must become equal to the Galilei transformation, i. e., $y' = y$, and therefore $L_{22} = +1$.

In the same way, we can analyze the xy-reversal and find $L_{33} = 1$. Then, only the coefficients L_{00}, L_{01} and L_{11} are left to determine.

Longitudinal direction: scale factor/inversion. In the next step, we analyze the transformation in the *direction of the relative velocity*. But first, we introduce the easier coefficients

$$\bar{\alpha}(v) := L_{00}(v) \ , \quad \bar{\beta}(v) := -L_{01}(v)/L_{00}(v) \ , \quad \bar{\gamma}(v) := L_{11}(v) \ ,$$

which brings us to

$$\begin{aligned} t' &= \bar{\alpha}(v)(t - \bar{\beta}(v)x) \ , \\ x' &= \bar{\gamma}(v)(x - vt) \ . \end{aligned} \tag{11.4}$$

To determine the remaining coefficients, the *inverse coordinate transformation* plays a central role (this is the transformation that Alice uses to calculate her coordinates (t, x) from Bob's coordinates (t', x') of an event). There are two ways to come to this transformation. The first one is just to invert (11.4), and the second one is via the *v-reversal*: replace $-v$ in (11.4) with v and the primed coordinates with the unprimed ones. Both, because of the relativity principle, must yield the same result. From that, the unknown coefficients are determined. We can ease the formulas a little bit by first considering a special case. Alice's location, given by $x = 0$, in Bob's coordinates, must be given by $x' = -vt'$. But setting $x = 0$ in (11.4) yields $x' = -(\bar{\gamma}/\bar{\alpha})vt'$, and therefore we need $\bar{\alpha}(v) = \bar{\gamma}(v)$ and (11.4) becomes

$$\begin{aligned} t' &= \bar{\gamma}(v)(t - \bar{\beta}(v)x) , \\ x' &= \bar{\gamma}(v)(x - vt) . \end{aligned} \tag{11.5}$$

Next, we carry out the program sketched above. First, we invert (11.5), which yields

$$\begin{aligned} t &= \frac{1}{\bar{\gamma}(v) \cdot (1 - v\bar{\beta}(v))} \left(t' + \bar{\beta}(v)x'\right) , \\ x &= \frac{1}{\bar{\gamma}(v) \cdot (1 - v\bar{\beta}(v))} \left(x' + vt'\right) . \end{aligned} \tag{11.6}$$

And second, we perform the *v-reversal*. This yields

$$\begin{aligned} t &= \bar{\gamma}(-v)(t' - \bar{\beta}(-v)x') , \\ x &= \bar{\gamma}(-v)(x' + vt') . \end{aligned} \tag{11.7}$$

Both transformations have to be equal. This is the case if

$$\begin{aligned} \bar{\beta}(-v) &= -\bar{\beta}(v) , \\ \bar{\gamma}(v)\bar{\gamma}(-v) &= \frac{1}{1 - v\bar{\beta}(v)} . \end{aligned}$$

By virtue of the *inversion symmetry*, the prefactor $\bar{\gamma}(v)$ must be an even function in v. If we exchange left and right by replacing $x \to -x$ and $x' \to -x'$, we also have to change the sign of the velocity: $v \to -v$. Then, the second formula of (11.4) becomes $x' = \bar{\gamma}(v) \cdot (x - vt)$. This must be invariant, and therefore we have the result that $\bar{\gamma}(-v) = \bar{\gamma}(v)$, i.e., $\bar{\gamma}(v)$ is an even function in v, and $\bar{\gamma}(v)\bar{\gamma}(-v) = \bar{\gamma}(v)^2$. Basically, this says that a rod that moves with velocity v relative to Alice experiences the same length contraction as a rod that moves with velocity $-v$. Length contraction does not depend on the direction in which the rod moves. It depends only on the magnitude of its velocity.

Therefore, we arrive at

$$\begin{aligned} t' &= \bar{\gamma} \cdot (t - \bar{\beta}(v)x) , \\ x' &= \bar{\gamma} \cdot (x - vt) , \\ \bar{\gamma}(v) &= \frac{1}{\sqrt{1 - v\bar{\beta}(v)}} . \end{aligned} \tag{11.8}$$

So far, the requirements made are also valid for the Galilei transformation. And we recognize that, for $\bar{\beta}(v) = 0$, our transformation indeed becomes the Galilei transformation. Therefore, the essential step to arrive at the Lorentz transformation is still missing.

Longitudinal direction: light cone. This essential step that makes the difference between the Galilei and Lorentz transformations is the requirement that light world lines map to light world lines: Light pulses must be transformed to light pulses. A light pulse emitted by Alice at $(x = 0, t = 0)$ is described by $x = \pm ct$. *The Lorentz transformation must map the trajectory $x = ct$ to the trajectory $x' = ct'$.The same holds for $x = -ct$ and $x' = -ct'$.*

Putting $x = ct$ in the transformation (11.8) above, we get

$$\begin{aligned} t' &= \bar{\gamma} \cdot (1 - c\bar{\beta}(v))t \ , \\ x' &= \bar{\gamma} \cdot (c - v)t \ , \end{aligned}$$

which must result in $x' = ct'$, and therefore we need $\bar{\beta} = v/c^2$, which gives us $\bar{\gamma}(v) = 1/\sqrt{1 - v^2/c^2}$. This leads us to the Lorentz transformation

$$\begin{aligned} t' &= \gamma_v \cdot \left(t - \frac{v}{c^2}x\right) \\ x' &= \gamma_v \cdot (x - vt) \\ \gamma_v &= \frac{1}{\sqrt{1 - (v/c)^2}} \ , \end{aligned}$$

where we replaced $\bar{\gamma}$ with the usual γ_v.

That's it! We can conclude:

The **Lorentz transformation** (in usual units) and for the standard configuration is given by

$$\begin{aligned} t' &= \gamma_v \cdot \left(t - \frac{v}{c^2}x\right) \ , \\ x' &= \gamma_v \cdot (x - vt) \ , \\ y' &= y \ , \quad z' = z \ , \end{aligned} \tag{11.9}$$

where

$$\gamma_v = \frac{1}{\sqrt{1 - v^2/c^2}} \ .$$

Again, notice the surprising symmetry between space and time. While the rotation in Euclidean space only "mixes" space dimensions, the Lorentz transformation "mixes" space and time. Hermann Minkowski expressed this circumstance with the following impressive words, directed to the 80th Assembly of German Natural Scientists and Physicians in 1908 [Minkowski08]:

> The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth, space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.

Exercise 50: Show that, for two arbitrary four-vectors $(a_0, \boldsymbol{a})$ and $(b_0, \boldsymbol{b})$, the "scalar product" $a_0 b_0 - \boldsymbol{ab}$ is invariant.

11.4 Digression: The Lorentz Transformation from Empirics

In a well-known paper [Robertson49], Robertson deduces the Lorentz transformations mostly from experimental results. Three cornerstone experiments are needed, the Michelson-Morley experiment, the Kennedy-Thorndyke experiment and the Ives-Stilwell experiment. We show the main steps of his derivation here.

Robertson assumes that there is *one* **preferred reference frame** Σ in which light in vacuum propagates rectilinarly and isotropically with a constant speed c. He introduces a coordinate system with time τ and space coordinates ξ, η, ζ, and the clocks are synchronized via the constancy of the speed of light: the master clock at time τ_0 sends a light pulse to another clock at a distance ℓ from the master clock and, upon arrival, is set to the time $\tau_0 + \ell/c$.

Next, he introduces a further reference frame S with coordinates t and x, y, z that is moving with a constant velocity $\boldsymbol{v}$ with $|\boldsymbol{v}| < c$, relative to Σ. We choose the x-axis to coincide with the ξ-axis. Hence, we have $\xi = v\tau$.

Robertson makes **no assumptions** as to the *speed of light* or the *laws of physics* in S and, in particular, does not assume the *relativity principle*. These properties must follow from the arguments at hand and the three cited experiments.

Starting with the most general linear transformation T, which maps (t, x, y, z) to (τ, ξ, η, ζ), Robertson uses a smart choice of the directions of the coordinate axes and some symmetry arguments and incorporates the velocity v, which leads to

$$\begin{aligned}
\tau &= at + dx, \\
\xi &= avt + bx, \\
\eta &= gy, \quad \zeta = gz,
\end{aligned}$$

where $a(v)$, $b(v)$, $d(v)$ and $g(v)$ are functions of v. The core of his paper is to determine these functions from the said experiments, along with the definition of simultaneity.

The **first step** is to incorporate **Einstein synchronization** in S. We send a light signal from the origin event $(x = 0, t = 0)$ to the location with x-coordinate $x_E > 0$, where it arrives at t_E and is reflected back to $x = 0$, where it arrives at $t = t_0$. Einstein synchronization then defines $t_E = t_0/2$.

The light world line starting at the origin event is given by $\xi = c\tau$ or

$$a(c - v)t = (b - dc)x \ .$$

The world line back to the t-axis, which intersects it at $t = t_0$, is given by $\tau - \tau_0 = -\xi$ (where τ_0 is determined by the intersection) or

$$a(c + v)(t - t_0) = -(b + dc)x \ .$$

The two world lines must intersect at $t = t_0/2$. Solving the two equations above for x, equating them and setting $t = t_0/2$ yields

$$\frac{a(c - v)}{b - dc} = \frac{a(c + v)}{b + dc}$$

and, eventually,

$$d = bv/c^2 \ .$$

Thus, the implementation of Einstein synchronization brings us to

$$\begin{aligned} \tau &= at + bvx/c^2 \ , \\ \xi &= avt + bx \ , \\ \eta &= gy \ , \quad \zeta = gz \ . \end{aligned} \tag{11.10}$$

In the **second step**, we incorporate the result of the **Michelson-Morley experiment**.

The time needed for a light pulse to travel from the origin event O to some point (ξ, η, ζ) at distance ℓ from the origin is given by

$$\tau = \frac{1}{c}\sqrt{\xi^2 + \eta^2 + \zeta^2} = \frac{\ell}{c}$$

or, after using the transformation (11.10) and solving for t,

$$\begin{aligned} t &= \frac{\ell}{ac}\sqrt{b^2\frac{x^2}{\ell^2} + g^2\gamma^2\frac{y^2 + z^2}{\ell}} \ , \\ &= \frac{\ell}{ac}\sqrt{b^2\cos^2\vartheta + g^2\gamma^2\sin^2\vartheta} \ , \end{aligned}$$

where ϑ is the angle between the x-axis and the direction of the light ray or the orientation of the Michelson-Morley interferometer,[3] and

[3] Note that the forward and backward velocities must be the same, because we imposed Einstein synchronization. Furthermore, without length contraction, we would have $b = g = 1$, and therefore the velocity would depend on the direction.

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

is our well-known γ-factor.

The experiment shows that, if the orientation of the interferometer is changed, the traveling time stays constant. Therefore, the expression above must be independent of ϑ. This requires

$$b = g\gamma$$

and leads us to

$$\begin{aligned}\tau &= at + \gamma g v x/c^2 \ ,\\ \xi &= avt + \gamma g x \ ,\\ \eta &= gy \ , \quad \zeta = gz \ .\end{aligned}$$

Physically, this says that a slab that is parallel to the ξ-axis and moves in the ξ-direction with velocity v for the observer in Σ is length-contracted by the factor $g\gamma$, while a slab that is perpendicular to the ξ-axis but moves with the same velocity in the ξ-direction is length-contracted by the factor g. The relative length contraction of these two directions is given by γ.[4]

The **third step** is about the result of the **Kennedy-Thorndike experiment**, which is basically equal to the Michelson-Morley experiment, with the exception that one of the interferometer arms is longer than the other one by $\Delta\ell$ and the light must cover this additional distance. The result of the experiment indicated that the time Δt needed for that is independent of v. If the distance lies in the x-direction, this means that

$$\Delta t = \frac{\Delta\ell}{c}\frac{b}{a} = \frac{\Delta\ell}{c}\frac{g\gamma}{a}$$

must be independent of v. For $v \to 0$, we must have $\Delta t = \Delta\ell/c$, and therefore $g\gamma/a = 1$ or

$$g\gamma = a \ .$$

This leads us to

$$\begin{aligned}\tau &= a\cdot(t + vx/c^2) \ ,\\ \xi &= a\cdot(vt + x) \ ,\\ \eta &= \frac{a}{\gamma}y \ , \quad \zeta = \frac{a}{\gamma}z \ .\end{aligned}$$

Remember that a and γ are functions of v.

[4] FitzGerald and Lorentz assumed $g = 1$ and $b = \gamma = 1/\sqrt{1-(v/c)^2}$, which is the *Lorentz-FitzGerald contraction.*

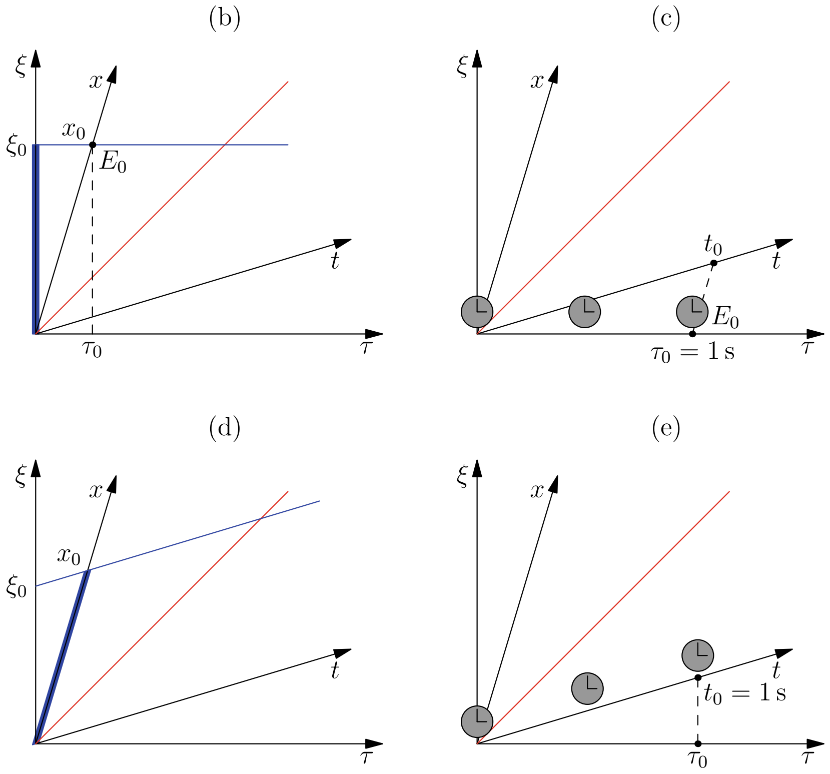


Fig. 11.3 To exercise 51

Physically, a is related to the time dilation and $g\gamma$ to the length contraction. Therefore, the Kennedy-Thorndike experiment fixes the relation between these two effects. The length of a faster moving object is contracted more than that of a slower object. To compensate, the time of the faster moving object is also dilated more than that of a slower object.

The only free parameter that is still to be determined is $a(v)$, the time dilation factor.

In the **fourth** and final **step**, we include the result of the **Ives-Stilwell experiment**, which has shown that the factor for the frequency change in the transversal Dopper effect, which is the same as the time dilation factor, has the value $\gamma(v)$. On the other hand, according to our transformation, this is $a(v)$. Therefore, $a = \gamma$, and we arrive at the *Lorentz transformation*

$$\begin{aligned}\tau &= \gamma(v)(t + vx/c^2)\ ,\\ \xi &= \gamma(v)(vt + x)\ ,\\ \eta &= y\ , \quad \zeta = z\ .\end{aligned}$$

The transformation for the transition from the preferred reference frame to another reference frame that moves uniformly relative to the former is therefore the **Lorentz transformation**. In particular, from the Lorentz transformation, it follows that

$$c^2\tau^2 - (\xi^2 + \eta^2 + \zeta^2) = c^2t^2 - (x^2 + y^2 + z^2) \ ,$$

which implies that light in vacuum propagates rectilinearly and isotropically with constant speed c in all reference frames that move uniformly relative to the preferred reference frame.

Applying the Lorentz transformation to **Maxwell's electrodynamics** shows us that this theory is **form-invariant**, it has the same form in both reference frames. Therefore, for this theory, the relativity principle holds.

Classical mechanics, however, is not form-invariant. Experiments show that mechanics must respect the relativity principle as well. Therefore, a new mechanics is needed that is form-invariant with respect to the Lorentz transformations and that becomes classical mechanics for small relative velocities. This is **special relativity**.

Exercise 51: Suppose there is a preferred reference frame Σ with coordinate system (τ, ξ) (we consider only $1 + 1$ dimensions). Furthermore, there is another coordinate system S with coordinates (t, x) whose axes are parallel to the related ones in Σ and whose origin in Σ moves according to $\xi = v\tau$.

(a) Show that the coordinate transformation is given by

$$\begin{aligned} t &= a(\tau - d\xi) \ , \\ x &= b(\xi - v\tau) \ , \end{aligned} \tag{11.11}$$

where a, b and d are functions of v. In special relativity, we have $a = b = \gamma(v)$ and $d = v/c^2$ (the Lorentz transformation), and in classical mechanics, we have $a = b = 1$ and $d = 0$ (the Galilei transformation).

(b) How does a moving observer experience a rod that is at rest in Σ? See Fig. 11.3, top left.

(c) How does a moving observer experience a clock that is at rest in Σ? See Fig. 11.3, top right.

(d) How does an observer at rest in Σ experience a rod that is at rest in S? See Fig. 11.3, bottom left.

(e) How does an observer at rest in Σ experience a clock that is at rest in S? See Fig. 11.3, bottom right.

What does the result demonstrate?

Chapter 12
The Lorentz Transformation: Applications

12.1 Again: The Effects of Special Relativity

In possession of the Lorentz transformation, we can derive the discussed relativistic effects easily.

Relativity of simultaneity. First, the relativity of simultaneity. For Alice, two events A and B are simultaneous if $t_A = t_B$. According to the Lorentz transformation, for Bob, $t'_A - t'_B = \gamma_v[(t_A - t_B) - (v/c^2)(x_A - x_B)] = -\gamma_v \cdot (v/c^2)(x_A - x_B)$ then holds. Provided that $v \neq 0$, this expression vanishes only if both events happen for Alice at the same place, but then A and B are the same event. Simultaneity for Alice is therefore different than simultaneity for Bob.

Time dilation. Then, time dilation (which is easier to demonstrate than length contraction). Consider Fig. 12.1, left side.[1] Alice and Bob carry clocks that have been synchronized and that reset when they meet in the origin. At event E_0, Bob's clock then shows the time t'_0, which is the clock's proper time. What does Alice's clock show? We draw a line of simultaneity (for Alice), which intersects Alice's t-axis at t_0. Because of the Lorentz transformation, $t_0 = \gamma_v(t'_0 + (v/c^2)x'_0)$. As for Bob's clock, $x'_0 = 0$, and thus we have $t_0 = \gamma_v t'_0$. Alternatively, using the time intervals starting at the origin, $\Delta t = \gamma_v \Delta t_0$. This is the expression for time dilation (see (9.1)).

Length contraction. Last, length contraction. We regard Fig. 12.1, right side. A rod with proper length l_0 moves with the velocity v relative to Alice and rests for Bob. The proper length is given by the distance from the origin to the event E_0 in Bob's coordinates. Alice measures the length l, given by the distance from the origin to the event E_1 in her coordinates (according to the rule for measuring lengths in Sect. 8.2.1, Alice must simultaneously determine the position of the front and rear ends of the rod, which corresponds to the distance from the origin to E_1).

[1] Note that our notation here differs from that in Sect. 9.2. This is meant to ease the comparison with length contraction.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_12

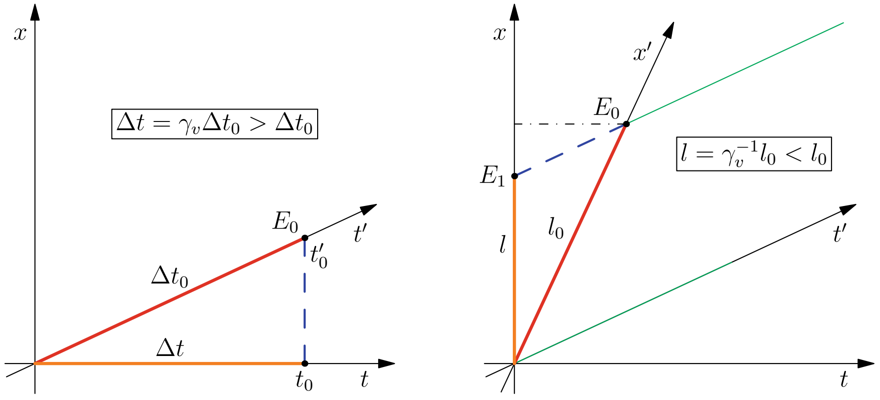


Fig. 12.1 To the derivation of time dilation (left side) and length contraction (right side) using the Lorentz transformation

We start the calculation with $E_1 = (0, l)$ in Alice's coordinates. E_0 then follows from $x = vt + l$ (green line through E_1) and Bob's time axis $x = (c^2/v)t$. This gives us $x_0 = \gamma_v^2 l$, $t_0 = (v/c^2)x_0$ and $x_0 - vt_0 = x_0\gamma_v^{-2} = l$. Plugging this into the Lorentz transformation, we get $x'_0 = \gamma_v(x_0 - vt_0) = \gamma_v l$. With $x'_0 = l_0$, it follows that $l = l_0/\gamma_v$, which is the expression (8.4) for length contraction.[2]

Comparison of length contraction and time dilation. You have surely already noticed that the result is not really symmetrical. Lengths become contracted, and thus shorter. The equation for this is $l = \gamma_v^{-1} l_0$ (l_0 is the *proper length* of the moving rod). Time intervals, however, are expanded, and thus longer. The equation for this is $\Delta t = \gamma_v \Delta t_0$ (Δt_0 is the *proper time* of the moving clock).[3] In one case, the factor γ_v appears, and in the other case, its inverse. What is the reason for this? Haven't we always stressed that space and time are symmetric in special relativity?

The reason for this is the different measurement rules. Look again at Fig. 12.1. In both cases, it is about an interval for Bob (in red in the figure): one is a time interval of length Δt_0 (the proper time indicated by the clock traveling with Bob) and one is a space interval of length l_0 (the proper length of the rod traveling with Bob).

In time dilation, *simultaneity for Alice* is relevant, and therefore a projection of E_0 along Alice's x-axis onto Alice's t-axis is performed (blue dashed line in the figure on the left side), yielding the time interval Δt (in orange in the figure). If the measurement rules were symmetrical, for the length contraction, *the same location* for Alice would be relevant and one would project E_0 along Alice's t-axis onto Alice's x-axis (indicated with a black dash-dotted line in the figure on the right side). But the measurement rules are different, and Alice performs the length measurement

[2] Note that, by plugging $x_0 - vt_0 = l$ into the Galilei transformation, $l = l_0$ follows—as expected.

[3] The "contraction" refers to the fact that the length of a moving object is shorter than it was when it rested. The "dilation" (see footnote 1 in Sect. 9.2) refers to the fact that the stationary clock runs faster than the moving one and, thus, the time on the moving clock is stretched.

of Bob's rod *simultaneously for herself*. This implies that she projects the event E_0 along Bob's t'-axis onto her x-axis (indicated with a blue dashed line in the figure on the right side). This yields the length l (in orange in the figure).

The terms time *dilation* and length *contraction* seem to imply that, in both cases, something becomes shorter. However, the perspective is different: in time dilation, *for Bob*, less time passes than for Alice, while in length contraction, *for Alice*, the rod is shorter than for Bob.

There's another source of confusion. In length contraction, there is only one rod, a rod that rests in Bob's frame and is the reference object. Therefore, length contraction says: "Bob's rod is shorter for Alice": $l = \gamma_v^{-1} l_0 < l_0$.

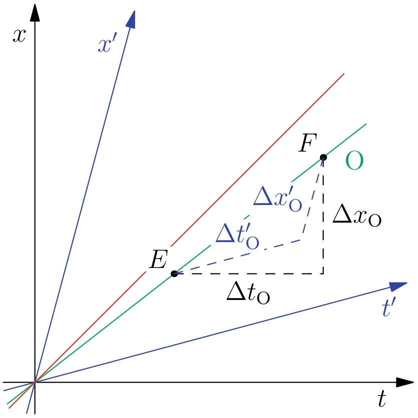
In time dilation, both Alice and Bob have a clock, and we can use either clock as a reference. If we choose Bob's clock as the reference clock, time dilation says: "When Bob's clock shows one second, Alice's clock will already show a later time. Alice's clock runs fast for Bob": $\Delta t_0 = \gamma_v \Delta t'_0 > \Delta t'_0$. If we choose Alice's clock as the reference, we have: "When Alice's clock shows one second, Bob's clock shows less than a second. Bob's clock runs slow for Alice": $\Delta t'_0 = \gamma_v^{-1} \Delta t_0 < \Delta t_0$. The conclusion, however, is always: Bob's clock runs slow for Alice (the situation is described from Alice's perspective).

A final point. If the Lorentz transformation were given by (11.4) with the actual equation $\bar{\beta} = v/c^2$ but $\bar{\alpha} \neq \bar{\gamma}$, then, according to our derivations of time dilation and length contraction above, the factor $\bar{\alpha}$ would determine the time dilation and the factor $\bar{\gamma}$ the length contraction.

Lorentzian addition of velocities. For the derivation of the Lorentzian addition of velocities, we consider our initial observers Alice and Bob with the unprimed and primed coordinate systems, respectively, in the standard configuration. Let v be Bob's velocity relative to Alice.

Furthermore, there is an object O that moves uniformly relative to Alice and Bob (see Fig. 12.2). If $x_\text{O}(t_\text{O})$ is the space trajectory of the object for Alice and $x'_\text{O}(t'_\text{O})$ that for Bob, the object has the velocity

Fig. 12.2 For the derivation of the formula for Lorentzian addition of velocities



$$u = \frac{\Delta x_\mathrm{O}}{\Delta t_\mathrm{O}} \quad \text{with } \Delta x_\mathrm{O} := x_F - x_E \text{ and } \Delta t_\mathrm{O} := t_F - t_E$$

for Alice and

$$u' = \frac{\Delta x'_\mathrm{O}}{\Delta t'_\mathrm{O}} \quad \text{with } \Delta x'_\mathrm{O} := x'_F - x'_E \text{ and } \Delta t'_\mathrm{O} := t'_F - t'_E$$

for Bob. Here, E and F are two different events on the object's trajectory.

The Galileian addition of velocities would then be $u' = u - v$. From the Lorentz transformation of the object's position, however, we get

$$\Delta x'_\mathrm{O} = \gamma_v(\Delta x_\mathrm{O} - v\Delta t_\mathrm{O}) = \gamma_v\left(\frac{\Delta x_\mathrm{O}}{\Delta t_\mathrm{O}} - v\right)\Delta t_\mathrm{O} = \gamma_v\,(u - v)\,\Delta t_\mathrm{O},$$
$$\Delta t'_\mathrm{O} = \gamma_v\left(\Delta t_\mathrm{O} - \frac{v}{c^2}\Delta x_\mathrm{O}\right) = \gamma_v\left(1 - \frac{v}{c^2}\frac{\Delta x_\mathrm{O}}{\Delta t_\mathrm{O}}\right)\Delta t_\mathrm{O} = \gamma_v\left(1 - \frac{uv}{c^2}\right)\Delta t_\mathrm{O}.$$

Through division, we immediately get

$$u' = \frac{u - v}{1 - \dfrac{uv}{c^2}}. \tag{12.1}$$

This is exactly the same as (77). We just have to replace $v_\mathrm{BA} \to v$, $v_\mathrm{CA} \to u$ and $v_\mathrm{CB} \to u'$ in (77) and resolve for u' (which can be done by exchanging u for u' and replacing v with $-v$ in the formula, see Exercise 41).

We note that the Lorentzian addition of velocities is nothing but the Lorentz transformation of velocities. And that this transformation is not as simple as that of space or time coordinates because, in the definition of the velocity, both, the nominator Δx and the denominator Δt transform.

In the following Exercise 52, we show an alternative derivation of the Lorentzian addition of velocities. And in the next section, Sect. 12.2, we will take a more in-depth look at the transformation behavior of the velocity.

Exercise 52: We deal with three inertial observers, Alice, Bob and Claire. As usual, Alice uses the non-primed, Bob the primed and Claire the double-primed coordinates (x'', t''). Bob moves with the velocity v relative to Alice and Claire with the velocity u relative to Bob.[4]

- Write down the Lorentz transformation $(x, t) \to (x', t')$ with the relative velocity v.
- Write down the Lorentz transformation $(x', t') \to (x'', t'')$ that is used to calculate Claire's coordinates from Bob's coordinates. Use the relative velocity v'.
- Express Claire's coordinates (x'', t'') according to those of Alice (x, t) and bring the formula to the usual form of the Lorentz transformation with relative velocity u. Show that the addition formula (77)

$$u = v \oplus v' = \frac{v + v'}{1 + \dfrac{vv'}{c^2}}$$

results and, at the same time,

$$\gamma_v \gamma_{v'} = \frac{\gamma_{v \oplus v'}}{1 + \dfrac{vv'}{c^2}} \tag{12.2}$$

holds.

- If you are still motivated, prove that, with (8.3), (12.2) indeed holds. This is required because, otherwise, performing one Lorentz transformation after another would not result in a Lorentz transformation. This is a longer calculation. Probably the shortest method is to take both sides of (12.2) to the power of (-2) and show that both sides are equal.

12.2 Digression: The Velocity Four-Vector

Comparison to the Lorentz transformation of space and time. The position x, together with the time t, transform in a Lorentz transformation into x' and t', and thus $(t, \boldsymbol{x})$ is a four-vector. Can we fit u into this scheme, i. e., can we find a "partner" for u such that, together, they transform as a four-vector? What would this partner be?

First of all, in $3 + 1$ dimensions, $\boldsymbol{x}$ and $\boldsymbol{u}$ are both (space) vectors, and therefore $\boldsymbol{u}$ would correspond to $\boldsymbol{x}$ in a four-vector. What is the accompanying quantity?

If we replace x with u in the Lorentz transformation, we get

$$u' = \gamma_v(u - vt). \qquad \text{(wrong attempt)}$$

So, u' is proportional to $u - vt$, but, according to (12.1), it should be proportional to $u - v$. Therefore, the partner of u is the constant 1. Replacing x with u and t with 1 in the Lorentz transformation (11.1) yields

$$\begin{aligned} u' &= \gamma_v \cdot (u - v), \\ 1 &= \gamma_v \cdot (1 - (v/c^2)u). \end{aligned} \qquad \text{(wrong attempt)}$$

Both equations are clearly wrong. Let us try to repair this defect by making the following replacement in the Lorentz transformation (11.1):

$$\begin{aligned} x &\to f(u) \cdot u, \\ t &\to g(u) \cdot 1. \end{aligned}$$

This yields

$$\begin{aligned} f(u')u' &= \gamma_v(f(u)u - vg(u)), \\ g(u') &= \gamma_v(g(u) - (v/c^2)f(u)). \end{aligned}$$

The first equation must become equal to (12.1). Therefore, we need $g(u) = f(u)$ and

$$f(u')u' = f(u)\gamma_v(u - v).$$

This, together with $u' = u \oplus (-v)$, determines $f(u)$.

Finding $f(u)$ is easier than was possibly expected. We can set $u = 0$, which implies that $u' = -v$, and we get $f(-v) \cdot (-v) = f(0)\gamma_v \cdot (-v)$, and then, taking $\gamma_{-v} = \gamma_v$ into account, $f(v) = f(0)\gamma_v$. For small velocities, $f(u)u$ must be equal to u, therefore, we need $f(0) = 1$, and hence $f(u) = g(u) = \gamma_u$.

The **four-vector of the velocity** eventually becomes

$$(\gamma_u c, \gamma_u u) \tag{12.3}$$

(remember that, in a four-vector, the first component is always the time component, while the second is the space component). The invariant (analogous to (9.2)) related to the velocity four-vector is just

$$\gamma_u^2(c^2 - u^2) = c^2.$$

The **Lorentz transformation for velocities** then reads as

$$\begin{aligned} \gamma_{u'}u' &= \gamma_v(\gamma_u u - (v/c)\gamma_u c), \\ \gamma_{u'}c &= \gamma_v(\gamma_u c - (v/c)\gamma_u u). \end{aligned} \tag{12.4}$$

For later use, we write these two equations in a slightly different form and call them **γ-formulas** (note that the second of these formulas equals (12.2)):

$$\begin{aligned} \gamma_{u'}u' &= \gamma_v\gamma_u(u - v), \\ \gamma_{u'} &= \gamma_v\gamma_u(1 - uv/c^2). \end{aligned} \tag{12.5}$$

The γ-formulas are nothing but the Lorentz transformation of velocities. And note again that dividing the first of these equations by the second one directly yields the Lorentzian addition of velocities.

Equation (12.4) tell us that the Lorentz transformation (11.9) of the velocity four-vector (12.3) corresponds exactly to the Lorentzian addition of velocities $u' = u \ominus v$: dividing the first equation of (12.4) by the second one yields $u' = (u - v)/(1 - uv/c^2)$.

Relation to the proper time. From (12.1), it becomes clear that the cumbersome denominator $(1 - uv/c^2)$ in the transformation of the velocity comes from the

Lorentz transformation of the denominator $\Delta t'$ in the expression $u' = \Delta x'/\Delta t'$ for the velocity. We could get rid of this problem by replacing $\Delta t'$ with another time difference that is invariant in a Lorentz transformation. And, indeed, we know such a quantity: it is the *proper time* from Sect. 9.10. Denote the proper time by τ and go from differences to differentials, after which we scrutinize the quantity

$$\frac{dx}{d\tau} = \frac{dx}{dt}\frac{dt}{d\tau} = u\gamma_u,$$

which is exactly the space component of the velocity four-vector, while $c\,dt/d\tau$ is the time component. Therefore, we can conclude that the velocity four-vector is given by

$$\left(c\frac{dt}{d\tau}, \frac{dx}{d\tau}\right) = \gamma_u(c, u).$$

Therefore, in relativity, we have two definitions of the velocity:

- The "usual" velocity $u = dx/dt$, which transforms according to the Lorentzian addition of velocities. We also say: it *transforms like a (relativistic) velocity* (which is different from how the position vector $\boldsymbol{x}$ transforms).
- The quantity $dx/d\tau = \gamma_u u$, which *transforms as the space-component of a four-vector.*

12.3 Digression: Addition of Non-parallel Velocities

To derive the formula for the addition of non-parallel velocities, we consider the exact same case as in Sect. 12.1, but this time, in three space dimensions.

Then, the object O, which moves uniformly relative to Alice and Bob, has the space trajectory $\boldsymbol{x}_{\mathrm{O}}(t_{\mathrm{O}})$ for Alice and $\boldsymbol{x}'_{\mathrm{O}}(t'_{\mathrm{O}})$ for Bob. The object's velocity is

$$\boldsymbol{u} = \frac{\Delta\boldsymbol{x}_{\mathrm{O}}}{\Delta t_{\mathrm{O}}} \quad \text{with } \Delta\boldsymbol{x}_{\mathrm{O}} := \boldsymbol{x}_F - \boldsymbol{x}_E \text{ and } \Delta t_{\mathrm{O}} := t_F - t_E$$

for Alice and

$$\boldsymbol{u}' = \frac{\Delta\boldsymbol{x}'_{\mathrm{O}}}{\Delta t'_{\mathrm{O}}}$$

for Bob.

The Galileian addition of velocities then would be

$$u'_x = u_x - v,$$
$$u'_y = u_y\,, \quad u'_z = u_z.$$

From the Lorentz transformation (11.9) of the object's position, however, we get

$$\Delta x'_{\mathrm{O}} = \gamma_v(\Delta x_{\mathrm{O}} - v\Delta t_{\mathrm{O}}) = \gamma_v \left(\frac{\Delta x_{\mathrm{O}}}{\Delta t_{\mathrm{O}}} - v\right) \Delta t_{\mathrm{O}} = \gamma_v \left(u_x - v\right) \Delta t_{\mathrm{O}},$$
$$\Delta y'_{\mathrm{O}} = \Delta y_{\mathrm{O}}\,, \quad \Delta z'_{\mathrm{O}} = \Delta z_{\mathrm{O}},$$
$$\Delta t'_{\mathrm{O}} = \gamma_v \left(\Delta t_{\mathrm{O}} - \frac{v}{c^2}\Delta x_{\mathrm{O}}\right) = \gamma_v \left(1 - \frac{v}{c^2}\frac{\Delta x_{\mathrm{O}}}{\Delta t_{\mathrm{O}}}\right) \Delta t_{\mathrm{O}} = \gamma_v \left(1 - \frac{u_x v}{c^2}\right) \Delta t_{\mathrm{O}}.$$

By division, we immediately arrive at the Lorentzian addition for non-parallel velocities

$$\begin{aligned} u'_x &= \frac{u_x - v}{1 - \dfrac{u_x v}{c^2}}, \\ u'_y &= \frac{1}{\gamma_v}\frac{1}{1 - \dfrac{u_x v}{c^2}} \cdot u_y, \\ u'_z &= \frac{1}{\gamma_v}\frac{1}{1 - \dfrac{u_x v}{c^2}} \cdot u_z, \end{aligned} \tag{12.6}$$

which corresponds to the addition $\boldsymbol{u}' = \boldsymbol{u} \oplus (-\boldsymbol{v})$, provided that the coordinate system is chosen such that $\boldsymbol{v}$ points in the x-direction.

An important observation here is that the γ-factor cancels out in the equation for u_x, but not in those for u_y and u_z (the transversal components). The denominator $1 - u_x v/c^2$ is always the same, because it originates from the transformation of the time.

Note also that the direction of the perpendicular component stays the same, because $u'_y/u'_z = u_y/u_z$, as required by the principle of relativity.

When both velocities are parallel, we have $u_x = u$ and $u_y = u_z = 0$, and get, from (12.6), the already known addition formula (12.1).

If $\boldsymbol{v}$ and $\boldsymbol{u}$, however, are orthogonal, we have $u_x = 0$, and therefore

$$\begin{aligned} u'_x &= -v, \\ u'_y &= \frac{u_y}{\gamma_v}, \\ u'_z &= \frac{u_z}{\gamma_v}. \end{aligned}$$

The reason for the factor γ_v in the denominator is that the velocities u_y, u_z were measured by Bob, and Bob's clocks, from Alice's point of view, run more slowly than her own clocks. Alice therefore thinks that Bob obtained a velocity of the observed object that is too large, and this is corrected by said factor.

Note that $\boldsymbol{u} \oplus \boldsymbol{v} \neq \boldsymbol{v} \oplus \boldsymbol{u}$ for non-parallel velocities. This is different from the Galilei transformation and is cause for some new effects. One of these is the *Thomas precession*.

Exercise 53: Show that, independent of v, from $|\boldsymbol{u}| = c$, it follows that $|\boldsymbol{u}'| = c$. The speed of light is the same in all inertial frames. Show also that $|\boldsymbol{u}| \leq c$ implies that $|\boldsymbol{u}'| \leq c$.

12.4 Relativistic Stellar Aberration

12.4.1 Including Relativistic Effects

The explanation of the stellar aberration in Sect. 4.4 has two defects. The first is that we assumed the existence of a luminiferous aether, which does not exist. The second is that we have used the Galilean addition of velocities. This is also not completely correct. We repair these defects now (it was not possible to perform this correction earlier, because we need the Lorentzian addition of non-parallel velocities to do so).

Abolishing the aether is easy. It just means that we can use whatever inertial frame we like to perform the calculation – including the inertial frame in which the Sun is at rest. Therefore, getting rid of the aether does not imply a change in the calculation. In the next section, we show that the aberration angle is indeed independent of the inertial frame used to perform the calculation.

So, what is left for us to change is to use the Lorentzian addition of velocities instead of the Galilean.

Consider again the situation in Fig. 4.20, left side. As in Sect. 4.4, $\boldsymbol{v}_{\text{RS}}$ is the velocity of the star relative to the Sun, $\boldsymbol{v}_{\text{ES}}$ the velocity of the Earth relative to the Sun, and, finally, $\boldsymbol{v}_{\text{RE}}$ the velocity of the star relative to the Earth.

As in Sect. 4.4, we use the coordinate system in such a way that

$$\boldsymbol{v}_{\text{ES}} = v_{\text{E}} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{v}_{\text{RS}} = -c \cdot \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix}, \tag{12.7}$$

where $v_{\text{E}} > 0$.

To become relativistically correct, instead of the Galileian addition of velocities $\boldsymbol{v}_{\text{RS}} = \boldsymbol{v}_{\text{RE}} + \boldsymbol{v}_{\text{ES}}$ (see (4.15)), we have to use the Lorentzian addition of velocities $\boldsymbol{v}_{\text{RS}} = \boldsymbol{v}_{\text{RE}} \oplus \boldsymbol{v}_{\text{ES}}$ or

$$\boldsymbol{v}_{\text{RE}} = \boldsymbol{v}_{\text{RS}} \oplus (-\boldsymbol{v}_{\text{ES}}).$$

Therefore, if we make the identification $\boldsymbol{v}_{\text{RE}} \to \boldsymbol{u}'$, $\boldsymbol{v}_{\text{RS}} \to \boldsymbol{u}$, and $\boldsymbol{v}_{\text{ES}} \to \boldsymbol{v}$, where $v \to -v_{\text{E}}$, we can directly use (12.6) and get

$$\begin{aligned} v_{\text{RE},x} &= \frac{v_{\text{RS},x} - v_{\text{E}}}{1 - v_{\text{RS},x} v_{\text{E}}/c^2}, \\ v_{\text{RE},y} &= \frac{1}{\gamma(v_{\text{E}})} \frac{v_{\text{RS},y}}{1 - v_{\text{RS},x} v_{\text{E}}/c^2}. \end{aligned}$$

Using (12.7) then yields

$$\begin{aligned} v_{\mathrm{RE},x} &= \frac{-c\cos\varphi - v_{\mathrm{E}}}{1 + (v_{\mathrm{E}}/c)\cos\varphi}, \\ v_{\mathrm{RE},y} &= \frac{-c\sin\varphi}{\gamma_{\mathrm{E}} \cdot (1 + (v_{\mathrm{E}}/c)\cos\varphi)}, \end{aligned} \tag{12.8}$$

where $\gamma_{\mathrm{E}} = \gamma(v_{\mathrm{E}}) = \gamma(-v_{\mathrm{E}})$.

With the definition of the angle φ', in (4.16) it finally results in the **relativistic aberration formula**

$$\tan\varphi' = \frac{v_{\mathrm{RE},y}}{v_{\mathrm{RE},x}} = \frac{\sin\varphi}{\gamma_{\mathrm{E}}(\cos\varphi + \beta_{\mathrm{E}})}. \tag{12.9}$$

The result, up to the relativistic factor γ_{E}, is the same as in (4.16). This factor, because of $v_{\mathrm{E}} \ll c$, deviates only slightly from one.

In his paper [Einstein05b], Einstein has given a formula that is different from (12.9), but equivalent to it. From $v_{\mathrm{RE},x} = -c\cos\varphi'$ and the first equation of (12.8), one gets

$$\cos\varphi' = \frac{\cos\varphi + \beta_{\mathrm{E}}}{1 + \beta_{\mathrm{E}}\cos\varphi}. \tag{12.10}$$

Still, one question has to be clarified: is it correct to apply the transformation law for velocities (Galileian or Lorentzian addition of velocities) to the velocity of waves? This question is subtle, the answer not so easy. In Sect. 4.5.3, we will see that the group velocity of a wave transforms like the velocity of an object, a (relativistic) velocity. The phase velocity, however, does not. In Sect. 12.5, then, we will show that this problem disappears completely in special relativity. But both the phase and the group velocity of *light waves* (in vacuum) transform as (relativistic) velocities.

12.4.2 Digression: Clarifications Regarding Stellar Aberration

Independence of the aberration angle δ from the inertial frame. We show now that the aberration angle δ is independent of the inertial frame used to calculate it. Suppose that standard conditions prevail and, in particular, Bob moves relative to Alice with the velocity $v = \beta c$ in the positive x-direction. If φ_{A} is the angle under which Alice sees a particular star, the angle φ_{B} under which Bob sees it is given by (12.10), i. e.,

$$\cos\varphi_{\mathrm{B}} = \frac{\cos\varphi_{\mathrm{A}} + \beta}{1 + \beta\cos\varphi_{\mathrm{A}}}. \tag{12.11}$$

To show that this formula is independent of the inertial frame, we introduce a further inertial frame S_0. Relative to this inertial frame, Alice and Bob move with the velocities $v_{\mathrm{A}0} = \beta_{A0}c$ and $v_{\mathrm{B}0} = \beta_{B0}c$, respectively. The coordinate axes x_0, y_0, z_0 of S_0

are parallel to the corresponding ones of Alice and Bob and the latter two move on the x_0-axis. If φ_0 is the angle under which the observer at rest in S_0 sees the star, we have

$$\cos\varphi_\mathrm{A} = \frac{\cos\varphi_0 + \beta_\mathrm{A0}}{1 + \beta_\mathrm{A0}\cos\varphi_0}, \tag{12.12}$$

$$\cos\varphi_\mathrm{B} = \frac{\cos\varphi_0 + \beta_\mathrm{B0}}{1 + \beta_\mathrm{B0}\cos\varphi_0}. \tag{12.13}$$

The independence of the description on the inertial frame is shown if these two formulas imply (12.11). To demonstrate this is rather easy, we just have to invert (12.12) and plug it into (12.13). Inversion is rather straightforward and gives us

$$\cos\varphi_0 = \frac{\cos\varphi_\mathrm{A} - \beta_\mathrm{A0}}{1 - \beta_\mathrm{A0}\cos\varphi_\mathrm{A}}.$$

From this, we get

$$\begin{aligned}
\cos\varphi_\mathrm{B} &= \frac{\cos\varphi_0 + \beta_\mathrm{B0}}{1 + \beta_\mathrm{B0}\cos\varphi_0} \\
&= \frac{(\cos\varphi_\mathrm{A} - \beta_\mathrm{A0}) + \beta_\mathrm{B0}(1 - \beta_\mathrm{A0}\cos\varphi_\mathrm{A})}{(1 - \beta_\mathrm{A0}\cos\varphi_\mathrm{A}) + \beta_\mathrm{B0}(\cos\varphi_\mathrm{A} - \beta_\mathrm{A0})} \\
&= \frac{\cos\varphi_\mathrm{A}(1 - \beta_\mathrm{A0}\beta_\mathrm{B0}) + (\beta_\mathrm{B0} - \beta_\mathrm{A0})}{(1 - \beta_\mathrm{A0}\beta_\mathrm{B0}) + (\beta_\mathrm{B0} - \beta_\mathrm{A0})\cos\varphi_\mathrm{A}} \\
&= \frac{\cos\varphi_\mathrm{A} + (\beta_\mathrm{B0} \ominus \beta_\mathrm{A0})}{1 + (\beta_\mathrm{B0} \ominus \beta_\mathrm{A0})\cos\varphi_\mathrm{A}}
\end{aligned}$$

with

$$\beta_\mathrm{B0} \ominus \beta_\mathrm{A0} = \frac{\beta_\mathrm{B0} - \beta_\mathrm{A0}}{1 - \beta_\mathrm{A0}\beta_\mathrm{B0}},$$

which is equal to β and which is what we had to show.

Wrong arguments. Since the inception of special relativity, there have been people who claim that stellar aberration proved it wrong. This incorrect claim is based on a false understanding of the physics of stellar aberration.

Remember that, when discussing the Doppler effect of light, we concluded that only the velocities of the source and the observer relative to the luminiferous aether matter. In special relativity, there is no longer any aether and, likewise, no velocities relative to the aether. Only the velocity of the observer relative to the source is left, and the Doppler effect only depends on this relative velocity.[5]

In a similar vein, some people claim that, according to special relativity, stellar aberration can only depend on the relative velocity of the observer and the star. Thus, they continue, this is not consistent with the experimental findings, because observations show that stellar aberration does not depend on the motion of a star.

Suppose that the velocity v_E in the formula (4.18) for the aberration angle δ in stellar aberration were the velocity of the Earth relative to the star (and not the relative velocity between two positions of the Earth, as is correctly the case). Now consider two different stars. The first is at rest relative to the Sun. Thus, the velocity of the Earth $v_{E,1}$ relative to this first star is equal to the velocity of the Earth relative to the Sun, and we can apply what we learned in Sect. 4.4.2. For the perspective of the Earth, we get an aberration ellipse with a semi-major axis of $v_E/c = 20.5''$. Let the second star move with a large velocity v_S relative to the Sun, much larger than the velocity of the Earth, relative to the Sun. Then, the velocity of the Earth relative to this second star would be $v_{E,2} \approx v_S$ and we would get an aberration angle v_S/c much larger than $20.5''$. This is not observed, as all stars perform ellipses with a semi-major axis of $20.5''$.

Indeed, special relativity is not wrong, but these arguments are. We have seen that aberration cannot depend on the relative velocity between the observer and the star because the direction of the light wave of the star's light at the location of the observer only depends on the location where the star was when the wave was emitted. The velocity of the star when the wave was emitted and the true location of the star when the wave was detected are completely irrelevant. Aberration is not an effect between a source and an observer, but rather between two observers.

What we have to show is that the aberration angle is independent of the inertial frame. We can calculate the observation angle φ_B of Bob from that of Alice and determine δ from these, or we can go to a different inertial frame, for instance, that in which the star is at rest, calculate φ_B and φ_A and determine δ from the latter, whereupon we must get the same result. This is what we showed at the beginning of this section.

[5] This statement has to be made with more precision. If a source emits a wave and the observer measures the frequency of this wave, then the observer's finding can only depend on the wave around the observer. This is in accordance with the principle of locality, which states that an object can be directly influenced only by its immediate surroundings, and is, on the other hand, required by the fact that information cannot be transmitted with a velocity faster than light. In the moment when the observer measures the frequency, the source could already no longer exist. Another possibility is to say that the Doppler effect can depend only on the difference of the velocity of the source when the wave was emitted and the velocity of the observer when the wave was observed, as we stated in Footnote 14. Note also that special relativity does not directly refer to relative velocities. Special relativity claims that Einstein's principle of relativity holds and that the Doppler effect cannot depend on the inertial frame used to describe (or measure) it. Therefore, only relative velocities can count, and these velocities can only be taken at events that are on the same light cone. The difference between two velocities "taken at the same time" would be observer-dependent.

12.5 Lorentz Transformation of Waves

In Sect. 9.6, we have seen that the frequency ν of a wave transforms differently in special relativity than in classical mechanics: the Doppler formula acquires an additional factor of γ_v. To describe completely the transformation of a wave, we still have to find out what happens to the wavevector $\boldsymbol{k}$.

12.5.1 *Invariance of the Phase and Transformation of a Wave*

As shown in Sect. 4.5, the Lorentz transformation must leave the wave's phase $\boldsymbol{kr} - \omega t$ invariant: the phase φ at an event E must be independent of the coordinate system. Otherwise, the phase of a wave would provide a means to distinguish between two inertial frames – in contradiction to the principle of relativity.

We start with **one space dimension**. If (t, x) and (t', x') are Alice's and Bob's coordinates of an arbitrary event E, respectively, and $\varphi(x, t)$ and $\varphi'(x', t')$ are the phase of the wave, the relation

$$\varphi(x, t) = \varphi'(x', t')$$

must be satisfied. In other words, we need

$$kx - \omega t = k'x' - \omega' t'$$

to be fulfilled.

Hence, for the Lorentz transformation

$$\begin{aligned} x &= \gamma_v \cdot \left(x' + vt'\right), \\ t &= \gamma_v \cdot \left(t' + \frac{v}{c^2}x'\right), \end{aligned}$$

we need

$$\begin{aligned} kx - \omega t &= k\gamma_v \cdot (x' + vt') - \omega\gamma_v \cdot \left(t' + \frac{v}{c^2}x'\right) \\ &= \gamma_v \cdot \left(k - \frac{v}{c^2}\omega\right)x' - \gamma_v(\omega - vk)t' \\ &\overset{!}{=} k'x' - \omega' t'. \end{aligned}$$

The direct consequence is that the wavevector k and the angular frequency ω, under Lorentz transformations, must transform as follows:

$$\begin{aligned} k' &= \gamma_v \cdot \left(k - \frac{v}{c^2}\omega\right), \\ \omega' &= \gamma_v \cdot (\omega - vk). \end{aligned} \tag{12.14}$$

This is the Lorentz transformation for k and ω.

If we go to **three space dimensions**, $kx - \omega t$ is replaced with $\boldsymbol{kx} - \omega t$ and, in the same way as above, we get (remember that the relative velocity v points in the x-direction)

$$\begin{aligned} k'_x &= \gamma_v \cdot \left(k_x - \frac{v}{c^2}\omega\right), \quad k'_y = k_y, \quad k'_z = k_z, \\ \omega' &= \gamma_v \cdot (\omega - vk_x). \end{aligned} \tag{12.15}$$

Just like the transversal components y and z of the position vector, the transversal components k_y and k_z of the wavevector do not transform in a Lorentz transformation.

If we replace $\boldsymbol{k}$ with $\boldsymbol{r}$ and ω/c with ct, this becomes the Lorentz transformation of space and time. Therefore, $(\omega/c, \boldsymbol{k})$ is a four-vector. It is called the **four-wavevector** or *wave four-vector* .

The relativistic invariant related to the four-wavevector is

$$\omega^2 - c^2\boldsymbol{k}^2. \tag{12.16}$$

For light in vacuum and in an inertial frame, we have $\omega = c|\boldsymbol{k}|$, and therefore $\omega^2 = c^2\boldsymbol{k}^2$. Then, according to (12.16), in any other inertial frame, we have $\omega'^2 = c^2\boldsymbol{k}'^2$ or $\omega' = c\left|\boldsymbol{k}'\right|$. This is nothing but a manifestation of the principle of the absolute speed of light. Note that, in the classical case (Galilei transformation), according to (4.22), we would have $\omega'(\boldsymbol{k}') = c|\boldsymbol{k}| - \boldsymbol{vk}$.

12.5.2 The Doppler Effect and Aberration for Light Waves

Now that we know how to Lorentz transform a general wavevector, we can easily derive the formulas for the Doppler effect and aberration.

We have already treated *aberration* several times. First in Sects. 4.2.5 and 4.4 under the assumption of a luminiferous aether and without taking the effects of special relativity into consideration. Then, in Sect. 12.4, this time relativistically. But in all these discussions, we, strictly speaking, considered velocities instead of waves. The *Doppler effect* was already treated classically in Sect. 4.2.4 and relativistically in Sect. 9.6. We repeat the discussions here, but in a different setting. This time, we make use of our knowledge on how the four-wavevector $(\omega/c, \boldsymbol{k})$ transforms. This implies a different setting. In the former sections, either Alice and Bob carried the source and the other was the observer. Now, we only refer to the wave that was emitted at some time by the source, and the observers Alice and Bob.

Let Alice and Bob be in standard configuration, i. e., Bob moves relative to Alice with the velocity $\boldsymbol{v} = v\boldsymbol{e}_x$.

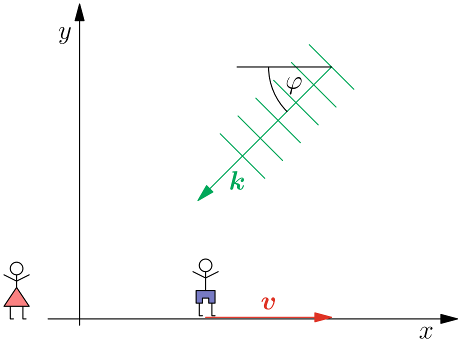


Fig. 12.3 For the derivation of the formulas for the relativistic Doppler effect and aberration

If $\boldsymbol{k}$ is the wave vector of the wave, $-\boldsymbol{k}$ points to the location where the source was when it emitted the wave. In the following, we use the unit vector $\boldsymbol{e}_{-\boldsymbol{k}} := -\boldsymbol{k}/k$, which points in the direction from which the wave comes.

Now let φ with $\cos\varphi = \boldsymbol{e}_{-\boldsymbol{k}}\boldsymbol{e}_x$ be the angle between $-\boldsymbol{k}$ and the direction of Bob's velocity relative to Alice, which is the x-direction. Then, the plane *light* wave moving with the phase velocity c has the four-wavevector

$$(\omega/c, k_x, k_y, k_z) = \frac{\omega}{c}(1, -\cos\varphi, -\sin\varphi, 0).$$

A Lorentz transformation (12.15) now yields the four-wavevector $(\omega'/c, k'_x, k'_y, k'_z) = (\omega'/c)(1, -\cos\varphi', -\sin\varphi', 0)$ for Bob:

$$\begin{aligned}
\omega' &= \gamma_v(\omega - vk_x) = \gamma_v\omega\left(1 + \frac{v}{c}\cos\varphi\right),\\
k'_x &= \gamma_v\left(k_x - \frac{v}{c^2}\omega\right) = -\gamma_v\frac{\omega}{c}\left(\cos\varphi + \frac{v}{c}\right) \overset{!}{=} -\frac{\omega'}{c}\cos\varphi',\\
k'_y &= k_y = -\frac{\omega}{c}\sin\varphi \overset{!}{=} -\frac{\omega'}{c}\sin\varphi',\\
k'_z &= 0\,.
\end{aligned}$$

The first equation is the **relativistic Doppler effect** for an arbitrary direction of motion of the source. If Bob travels in the exact direction from which the wave comes, we have $\boldsymbol{e}_v = \boldsymbol{e}_{-\boldsymbol{k}}$ and the formula becomes that for the longitudinal Doppler effect (9.4). If, however, Bob travels exactly perpendicular to the direction from which the wave comes (for Alice), we have $\boldsymbol{e}_v \perp \boldsymbol{e}_{-\boldsymbol{k}}$ and recover the formula for the transversal Doppler effect (9.5).

The remaining equations contain the relativistic aberration. Dividing these yields the relativistic aberration formula

$$\tan\varphi' = \frac{1}{\gamma_v}\frac{\sin\varphi}{\cos\varphi + \beta}$$

with $\beta = v/c$, which is equal to (12.9).

A word of caution. In one space dimension, one has to pay attention to the signs. Let us choose $v > 0$ and consider the second equation of (12.14). If the wave comes from Bob's right, then $k = -\omega/c$ has to be used, because the wave with $k > 0$ does not arrive at Bob's location. In this case,

$$\omega' = \gamma_v \omega \cdot (1 + v/c) > \omega \tag{12.17}$$

and the wave frequency is larger for Bob than for Alice. If, however, the wave comes from Bob's left then $k = +\omega/c$. In this case,

$$\omega' = \gamma_v \omega \cdot (1 - v/c) < \omega \tag{12.18}$$

and the wave frequency is smaller for Bob.

12.5.3 The Transformation of the Wavevector in Classical and Relativistic Physics

Let us compare the Galilei transformation (left side, see also (4.20)) with the Lorentz transformation (right side, see also (12.14)) of the angular frequency ω and the wave number k of a wave:

$$\begin{aligned} \omega' &= \omega - vk, & \qquad \omega' &= \gamma_v \cdot (\omega - vk), \\ k' &= k, & \qquad k' &= \gamma_v \cdot \left(k - \frac{v}{c^2}\omega\right). \end{aligned}$$

The transformation of the*frequency* ω is basically the Doppler formula. For example, if we deal with a wave with phase velocity c_{W} in the unprimed inertial frame, we just have to set $\omega = c_{\mathrm{W}} k$ in the transformation formulas for ω (this means that there is a medium and it is at rest in the unprimed reference frame) and get the classical Doppler formula $\omega' = \omega \cdot (1 - v/c_{\mathrm{W}})$ or the relativistic one $\omega' = \omega \cdot \gamma_v \cdot (1 - v/c_{\mathrm{W}})$, the difference between the two being merely just the additional factor γ_v.

For the *wavevector* k, the situation is very different. In the Galilei transformation, it does not transform at all, while, in the Lorentz transformation, it transforms like the space component of a four-vector. The ultimate reason for this is that time is absolute in the Galilei and relative in the Lorentz transformation.

Why? Consider the situation in three space dimensions. The wave vector $\boldsymbol{k}$ is perpendicular to the wavefront and points in the direction of an increasing phase (for fixed time). The length of $\boldsymbol{k}$ is given by 2π divided by the wavelength λ. The latter is the minimal distance between two wavefronts to the same phase.

A wavefront (which is a plane for plane waves) is the set of all points that have the same phase *at the same time*. Different concepts of simultaneity lead to different wavefronts. This is the reason why $\boldsymbol{k}$ is absolute in classical physics and relative in special relativity.

We will now demonstrate geometrically that a wavefront is different for different inertial observers.

Why the wavevector transforms. To this end, we need at least two space dimensions, and therefore our explanation uses (2+1)-dimensional spacetime.[6] In two space dimensions, wavefronts are lines (instead of planes, as in three space dimensions) and a wavefront in (2+1)-dimensional spacetime traces a two-dimensional *world plane*[7] (in the same way as an event traces a world line).

Suppose that a plane *light* wave ($\omega = ck$) moves in the positive y-direction (see Fig. 12.4).[8] The figure on the left shows the world plane of the wavefront given by $y = 0, t = 0$ in red. The world plane of the wavefront is said wavefront's trajectory. The spacial form of the wavefront at a fixed time t is the line that results from intersecting the world plane with a plane of fixed time t. And due to the fact that the planes of simultaneity are different for different inertial observers, the spacial form of the wavefront is also different for different inertial observers. This is what it means when we say that the wave number vector $\boldsymbol{k}$ transforms.

Now, what do the wavefronts look like for our inertial observers? From Alice's point of view, we intersect the world plane of the wavefront with a plane of fixed t, and therefore with a plane parallel to the x-y-plane, and get one of the blue lines in Fig. 12.4 on the left and the middle. For Alice, the wave vector $\boldsymbol{k}$ is perpendicular to it, and therefore parallel to the y-axis: $\boldsymbol{k} = (0, k_y)$.

Bob moves relative to Alice with the velocity v in the direction of the *negative* x-axis, $\boldsymbol{v} = (-v, 0)$, which is perpendicular to the wavevector $\boldsymbol{k}$ (for Alice). Bob's coordinate system has the coordinates x' and y', which are related to Alice's coordinates via a Lorentz transformation (11.9) (with v replaced with $-v$). To determine the world line of the wavefront for Bob, we again have to take the world plane of the wavefront for a fixed time t' (i. e., for Bob) and intersect it with a plane parallel to the x'-y'-plane. The lines that we get in this way are depicted in green in the figure on the left and on the right and are not parallel to the x'-axis. For this reason, the wavevector $\boldsymbol{k}'$ is not parallel to the y'-axis: $\boldsymbol{k}' = (k'_x, k'_y)$ with $k'_x \neq 0$.

Determination of the change of $\boldsymbol{k}$ from the spacetime transformation. Let us geometrically determine this effect from the Lorentz transformation of spacetime and Fig. 12.4. The Lorentz transformation is given by

[6] For an extensive discussion of aberration and the transformation of the wavevector, see, for instance, [LiebscherBrosche98].

[7] This term is non-standard. Sometimes, the term *world sheet* is used for the surfaces that strings are tracing in spacetime. Here, we have a special string: a line.

[8] Actually, there is no reason to restrict these considerations to a light wave. The argument works equally well for waves with a phase velocity $v_\mathrm{p} < c$.

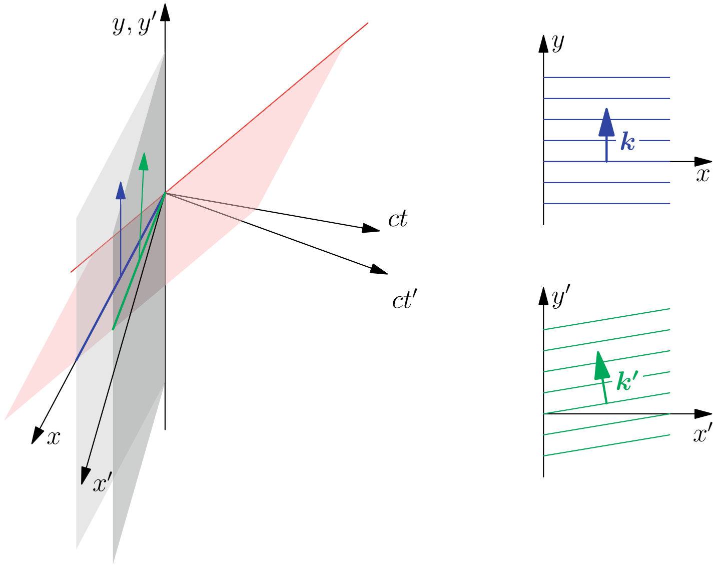


Fig. 12.4 Transformation of a wavefront or a wavevector $\boldsymbol{k}$, respectively

$$x' = \gamma_v \cdot (x + vt), \quad y' = y,$$
$$t' = \gamma_v \cdot \left(t + \frac{v}{c^2}x\right)$$

and the wavefront by $y = ct$. What do the wavefronts in space quantitatively look like?

For Alice, at $t = 0$, the wavefront is simply given by $y = 0$. The wave vector $\boldsymbol{k}$ is proportional to the unit vector $\boldsymbol{e}_k = (0, 1)$. For Bob, we take $t' = 0$. From $y' = y = ct = c\gamma_v \cdot (t' - (v/c^2)x')$ follows $y' = -\gamma_v \cdot (v/c)x'$ for the wavefront at $t' = 0$. The wavevector $\boldsymbol{k}$ must be perpendicular to the wavefront, and therefore is proportional to $(\gamma_v \cdot v/c, 1)$, i. e., parallel to the unit vector $\boldsymbol{e}_{k'} = (v/c, \gamma_v^{-1})$. Therefore, *the wavefronts* (or the wavevector) *rotate* (or are sheared, to be more precise). The reason for this is the relativity of simultaneity.

Direct Lorentz transformation of the four-wavevector. Let us directly transform the four-wavevector, which is $\boldsymbol{k} = k_0 \cdot (0, 1)$ and $\omega = ck_0$ for Alice. For Bob, we get

$$\begin{aligned} k'_x &= \gamma_v \cdot \left(k_x + \frac{v}{c^2}\omega\right) = \gamma_v \cdot \frac{v}{c}k_0, \\ k'_y &= k_y = k_0, \\ \omega' &= \gamma_v \cdot (\omega + vk_x) = \gamma_v \omega = \gamma_v \cdot ck_0. \end{aligned}$$

Therefore, the wavevector is $\boldsymbol{k}' = k_0 \cdot (\gamma_v \cdot v/c, 1)$ for Bob. The associated unit vector is $\boldsymbol{e}_{k'} = (v/c, \gamma_v^{-1})$, exactly as determined above geometrically.

For the magnitude of the wavevector in the primed reference frame, we get

$$\left|\boldsymbol{k}'\right| = k_0 \cdot \sqrt{1 + \gamma_v^2 \frac{v^2}{c^2}} = \gamma_v k_0,$$

which is larger than k_0. This is clear because the y-component stays the same and a non-vanishing x-component appears. Note also that $\omega' = c\left|\boldsymbol{k}'\right|$, which is required by the principle of the absolute speed of light.

From this, for the wavelength, we get

$$\lambda' = \lambda_0/\gamma_v,$$

and therefore the wavelength is smaller for Bob. Note that this is not directly explainable with length contraction. As the relative motion of Alice and Bob is in the x-direction, there is only a length contraction in the x-direction. However, the wave, as seen by Alice, travels in the y-direction. The fact that, for Bob, the x-component does not vanish causes the wavelength to shrink, and the shrinking factor is γ_v due to an interplay between both coordinate directions.

12.5.4 Again: The Velocity of a Wave

In this section, we investigate how the phase and group velocity of a wave transform. We consider waves in general, not only light waves.

Light waves. To discuss (12.14), we first consider light waves (waves with a phase velocity of c). For these, $\omega = ck$ and the Lorentz transformation of the *angular frequency* becomes nothing but the relativistic Doppler effect that we discussed in Sect. 9.6 (here, for $v > 0$, source and observer move toward each other) and demonstrated in Sect. 12.5.3. Remember that the difference between the classical and the relativistic Doppler effect, the factor γ_v, is due to time dilation. The transformation of the *wave number* becomes

$$k' = \gamma_v \cdot \left(1 - \frac{v}{c}\right) k,$$

which is the same formula again as for the relativistic Doppler effect when $\omega' = ck'$. But the latter relation must be satisfied, as the speed of light is the same in all inertial frames. Hence, for light, the transformation of k is not very surprising.

We show now that the *phase velocity*, the *group velocity*, and the unit vector $\boldsymbol{e}_k := \boldsymbol{k}/k$ of the *wavevector* are all equal for a wave with speed c and that this is true for all inertial observers.

As we have shown already (for instance, in (12.16)), for light, the dispersion relation in the unprimed reference frame is $\omega_{\mathbf{k}} = c|\mathbf{k}|$, while, in the primed reference frame, it is $\omega'_{\mathbf{k}'} = c\left|\mathbf{k}'\right|$. Therefore, it has the same form in both inertial frames, and this is required because of the principle of the absolute speed of light.

The *phase velocity* $\boldsymbol{v}_\mathrm{p}$ and the *group velocity* $\boldsymbol{v}_\mathrm{g}$ are defined by (4.23) and (4.24), respectively. They are equal for dispersion relations of the form $\omega_\mathbf{k} = v_\mathrm{p}|\mathbf{k}|$ and we have $\boldsymbol{v}_\mathrm{g} = c\boldsymbol{e}_k$ and $\boldsymbol{v}_\mathrm{g} = c\boldsymbol{e}_{k'}$ for light. Both are directly proportional to the unit vector $\boldsymbol{e}_k$ (or $\boldsymbol{e}_{k'}$) of the wave number. To show that all three quantities transform as (relativistic) velocities, we have to prove this only for one of them, and we select c times $\boldsymbol{e}_k$, the unit vector of the wavevector $\mathbf{k}$.

Let us choose the standard configuration, so that the relative velocity between the two inertial frames $\boldsymbol{v}$ points in the positive x-direction. Furthermore, if we denote the components of the phase velocity by $v_{\mathrm{p},i}$, where $i = x, y, z$, we have

$$v_{\mathrm{p},i} = c\frac{k_i}{k},$$

where k_i is the i-component of the wavevector and $k := |\boldsymbol{k}|$ its magnitude. We deduce the transformation behavior of the $v_{\mathrm{p},i}$ from that of the k_i and k.

The quantity $(\omega/c, \boldsymbol{k})$ is a four-vector, which means that it transforms as in (12.15). For light, we have $\omega_\mathbf{k} = c|\mathbf{k}| = ck$ and $\omega'_{\mathbf{k}'} = c\left|\mathbf{k}'\right| = ck'$, and therefore

$$k' = \gamma_v \cdot \left(k - \frac{v}{c}k_x\right) = \gamma_v \cdot \left(1 - \frac{v}{c}\frac{k_x}{k}\right)k = \gamma_v \cdot \left(1 - \frac{v v_{\mathrm{p},x}}{c^2}\right)k$$

from the Lorentz transformation of ω. On the other hand, the Lorentz transformation of k_x gives us

$$k'_x = \gamma_v \cdot \left(k_x - \frac{v}{c}k\right) = \gamma_v \cdot \left(\frac{v_{\mathrm{p},x} - v}{c}\right)k.$$

The components of the phase velocity $\boldsymbol{v}_\mathrm{p}$ (or wavevector's unit vector $\boldsymbol{e}_k$) therefore transform as

$$\begin{aligned}
v_{\mathrm{p},x} &= c\frac{k'_x}{k'} = \frac{\gamma_v \cdot (v_{\mathrm{p},x} - v)k}{\gamma_v \cdot \left(\frac{1 - v v_{\mathrm{p},x}}{c^2}\right)k} = \frac{v_{\mathrm{p},x} - v}{1 - \frac{v v_{\mathrm{p},x}}{c^2}},\\
v_{\mathrm{p},y} &= c\frac{k'_y}{k'} = \frac{k_y}{\gamma_v \cdot \left(\frac{1 - v v_{\mathrm{p},x}}{c^2}\right)k} = \frac{1}{\gamma_v}\frac{1}{1 - \frac{v v_{\mathrm{p},x}}{c^2}} \cdot v_{\mathrm{p},y},\\
v_{\mathrm{p},z} &= c\frac{k'_z}{k'} = \frac{k_z}{\gamma_v \cdot \left(\frac{1 - v v_{\mathrm{p},x}}{c^2}\right)k} = \frac{1}{\gamma_v}\frac{1}{1 - \frac{v v_{\mathrm{p},x}}{c^2}} \cdot v_{\mathrm{p},z}.
\end{aligned}$$

Comparison with (12.6) shows that the phase velocity transforms exactly as a (relativistic) velocity. Therefore, the phase velocity of light waves is a (relativistic) velocity, as are the group velocity and the unit vector of the wavevector.

We summarize: for light waves,

$$\boldsymbol{v}_\mathrm{g} = \boldsymbol{v}_\mathrm{p} = c\boldsymbol{e}_k$$

and all these quantities **transform as (relativistic) velocities**.

"Slow" waves. Second, we look at slow waves with $\omega = uk$ and $u < c$. In this case, (12.14) yields

$$k' = \gamma_v \cdot \left(k - \frac{v}{c^2}\omega\right) = \gamma_v \cdot \left(1 - \frac{uv}{c^2}\right) k,$$
$$\omega' = \gamma_v \cdot (\omega - vk) = \gamma_v \cdot (u - v)k,$$

and therefore

$$\frac{\omega'}{k'} = \frac{u - v}{1 - \frac{uv}{c^2}}.$$

Because $v_\mathrm{p} = \omega/k = u$ is the phase velocity, this means that the phase velocity of a wave transforms like a relativistic velocity. This, however, is only true if the phase velocity is parallel to the relative velocity $\boldsymbol{v}$ of the observers. Otherwise, for "slow" velocities, the phase velocity no longer transforms like a (relativistic) velocity. In other words: the phase velocity of "slow" waves is no longer a (relativistic) velocity. The group velocity of "slow" waves, however, still transforms as a velocity. Therefore, when we talk about the velocity of a "slow" wave, we should have its group velocity in mind.

We summarize: for "slow" waves, in general,

$$\boldsymbol{v}_\mathrm{g} \neq \boldsymbol{v}_\mathrm{p} \ .$$

The group velocity $\boldsymbol{v}_\mathrm{g}$ **transforms as a (relativistic) velocity**. The phase velocity $\boldsymbol{v}_\mathrm{p}$ transform as a (relativistic) velocity **only if** it is parallel to the relative velocity $\boldsymbol{v}$ of the observers.

Note that the fact that the phase and the group velocity transform differently for $u < c$ is not a contradiction of the principle of relativity. For waves with $u < c$, a medium is always needed, and the rest frame of this medium is a special reference frame. If we put the observer in a different inertial frame, but not the medium, the principle of relativity does not apply.

12.6 The Michelson-Morley Experiment Revisited

In Sect. 5.2, we tried to describe the Michelson-Morley experiment under the assumptions (a) that there is a **special inertial frame** (the "aether") in which the light moves with the speed c in all directions and (b) that the speed of light in other inertial frames is given by the Galileian addition of velocities.

Then, if the interferometer moves with the velocity v_sif relative to the supposed special inertial frame, for the light pulse in the interferometer arm, we get a traveling time (5.2) for the interferometer arm *parallel* to the direction of movement. This is different from (5.3), the traveling time for the interferometer arm *perpendicular* to the direction of movement. The experiment, however, shows that this is wrong, and one actually gets the same traveling time for both interferometer arms (provided that both have the same rest length).

What happens according to **special relativity**? There is no special inertial frame. We consider the following situation: both Alice and Bob are in an inertial frame. Bob's inertial frame moves with velocity v relative to Alice's inertial frame. And Bob carries the Michelson-Morley interferometer with him (see Fig. 12.5 on the left).

We describe the experiment first from Bob's and then from Alice's perspective. The lengths and time differences measured by Alice are denoted by L and T, respectively, and those measured by Bob are marked with a subscript 0, because these are the quantities measured at rest: L_0 is the rest length of the interferometer arms, which is considered to be the same for both arms.

Bob's perspective. Bob's description of the experiment is trivial. For him, the two interferometer arms are equal in length and the speed of light is absolute and equal in all directions. Therefore, the total traveling time for the light pulse (from the beam splitter to the mirror and back) is the same in both interferometer arms: $T_0 = 2L_0/c$. For the speed of light, Bob gets $2L_0/T_0 = c$.

Alice's description. For Alice, light moves with speed c in her inertial frame, and she uses her clock and her meter stick to measure times and lengths. We use the following times and lengths in the discussion:

$L_{\parallel}$	length of the parallel interferometer arm
$T_{\parallel +}$	time needed from beam splitter to mirror $M_{\parallel}$
$T_{\parallel -}$	time needed from mirror $M_{\parallel}$ to beam splitter
$T_{\parallel}$	time needed from beam splitter to mirror $M_{\parallel}$ and back
$L_{\perp}$	length of the perpendicular interferometer arm
$T_{\perp +} = T_{\perp -}$	time needed from beam splitter to mirror $M_{\perp}$ or back
$T_{\perp}$	time needed from beam splitter to mirror $M_{\perp}$ and back

and have

$$\begin{aligned} T_{\parallel} &= T_{\parallel +} + T_{\parallel -}, \\ T_{\perp} &= T_{\perp +} + T_{\perp -} = 2T_{\perp +} = 2T_{\perp -}. \end{aligned}$$

In (5.2) and (5.3), we calculated the time that light pulses need to pass through the interferometer arm *parallel* to the traveling direction of the interferometer and *perpendicular* to it, respectively. There, we assumed that the interferometer moves relative to a special inertial frame in which the speed of light is the same in all directions. Here, the interferometer moves relative to Alice's inertial frame and, for Alice, of course, the principle of the absolute speed of light holds. Therefore, the results (5.2) and (5.3) for the traveling time in the parallel and the perpendicular arm,

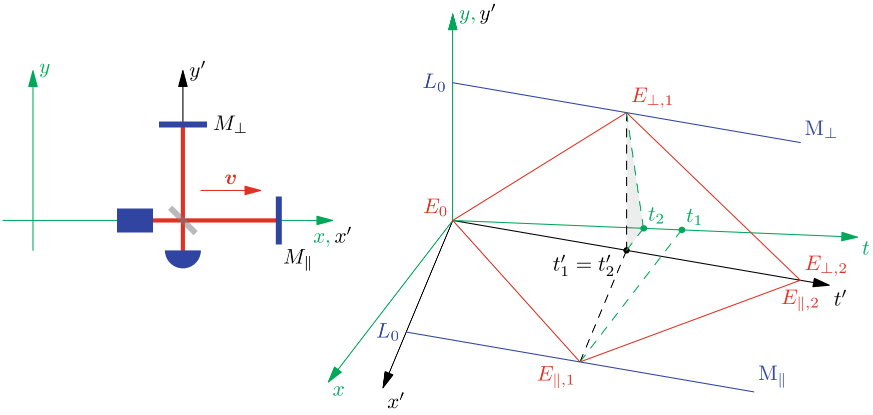


Fig. 12.5 A moving Michelson-Morley interferometer, from Alice's perspective (green coordinate axes with coordinates x, y, t). Left: In space. Right: In spacetime. The beam splitter is located on the t'-axis and $\mathrm{M}_\parallel$ and $\mathrm{M}_\perp$ denote the locations of the mirrors. The gray triangle is perpendicular to the x-y-plane

respectively, of the interferometer still hold. We just have to replace the velocity of the interferometer relative to the special inertial frame with the velocity relative to Alice.

There is one difference, however. The length of the parallel interferometer arm, for her, is contracted, $L_\parallel = L_0/\gamma_v$. From (5.2), she gets the total traveling time $T_\parallel = (2L_\parallel/c)\gamma_v^2 = 2\gamma_v L_0/c$. For the arm *perpendicular* to the traveling direction, she can directly use (5.3). The length of the interferometer arm is not contracted for her and she gets $T_\perp = 2\gamma_v L_0/c$. So, Alice gets equal traveling times $T_\parallel = T_\perp = 2\gamma_v L_0/c$. From her perspective, the effect of the *moving interferometer arms* on the traveling times of the light pulse is compensated by the *Lorentz contraction* of the interferometer arm lengths.

Indeed, length contraction is sufficient to explain the results of the experiment by Michelson and Morley, and this is what was done by FitzGerald and Lorentz (see Sect. 5.3.4). To demonstrate time dilation, other experiments are needed.

Light trajectories. We now follow the path of a light pulse in the Michelson-Morley interferometer (see Fig. 12.5, right side). The light pulse comes from the light source and gets split into two partial pulses at event E_0 in the beam splitter. First, we follow the partial pulse through the parallel interferometer arm. It gets reflected at the mirror at event $E_{\parallel,1}$ and is back to the beam splitter at $E_{\parallel,2}$, where it interferes with the other partial pulse. Second, we follow the other partial pulse through the perpendicular interferometer arm. It gets reflected back at the mirror at event $E_{\perp,1}$ and is back to the beam splitter at $E_{\perp,2}$, where it interferes with the other partial pulse. The events $E_{\parallel,2}$ and $E_{\perp,2}$ must be equal (same location, same time), otherwise, there would be no constructive interference. This is true for Bob, and therefore, it must also be true for Alice.

Let the interferometer arms have the rest length L_0. We denote Bob's coordinates with a prime and Alice's coordinates without a prime and use the standard configuration. Then, the Lorentz transformation is

$$\begin{aligned} t &= \gamma_v \cdot \left(t' + \frac{v}{c^2}x'\right), \\ x &= \gamma_v \cdot (x' + vt'), \\ y &= y' \end{aligned}$$

with $v > 0$. From this, we get, for the events:

Event	Bob (t', x', y')	Alice (t, x, y)
E_0	$(0, 0, 0)$	$(0, 0, 0)$
$E_{\parallel,1}$	$(1, c, 0) \cdot L_0/c$	$(1, c, 0) \cdot \gamma_v \cdot (1 + v/c)L_0/c$
$E_{\parallel,2}$	$(1, 0, 0) \cdot 2L_0/c$	$(1, v, 0) \cdot 2\gamma_v L_0/c$
$E_{\perp,1}$	$(1, 0, c) \cdot L_0/c$	$(\gamma_v, \gamma_v v, c) \cdot L_0/c$
$E_{\perp,2}$	$(1, 0, 0) \cdot 2L_0/c$	$(1, v, 0) \cdot 2\gamma_v L_0/c$.

Note that $E_{\parallel,2} = E_{\perp,2}$; we will also refer to this event as E_2.

Directly from the Lorentz transformation, for Alice, in the traveling times in parallel direction are

$$\begin{aligned} T_{\parallel+} &= \gamma_v \cdot (1 + v/c)L_0/c, \\ T_{\parallel-} &= T_\parallel - T_{\parallel+} = 2\gamma_v L_0/c - \gamma_v \cdot (1 + v/c)L_0/c = \gamma_v \cdot (1 - v/c)L_0/c, \\ T_\perp &= \gamma_v L_0/c, \end{aligned}$$

and with the contracted length $L_\parallel = L_0/\gamma_v$ and $L_\perp = L_0$, we can write

$$\begin{aligned} T_{\parallel\pm} &= \gamma_v \cdot (1 \pm v/c)L_0/c = \gamma_v^2 \cdot (1 \pm v/c)L_\parallel/c = \frac{1}{c \mp v}L_\parallel, \\ T_\perp &= \gamma_v \frac{L_\perp}{c}, \end{aligned}$$

which is exactly the same as the classical result (see (5.2) and (5.3)). Therefore, the γ_v-factors that appear here for Alice are not related to time dilation. For Alice, the interferometer moves and the light pulse has to travel a distance that is longer than L_0. In the perpendicular direction, the distance that the light travels from E_0 to E_2 is $2\gamma_v L_0 = 2\gamma_v^2 L_\parallel$. In the parallel direction, the distance must be the same for Alice and Bob, otherwise, the light pulses would not meet at E_2.

Interference pattern. In the Michelson-Morley experiment, the experimenter actually does not measure the traveling time of light pulses in the two interferometer arms, but rather observes an **interference pattern**. And this interference pattern must not depend on the motion of the Michelson-Morley interferometer, i. e., it must not depend on the inertial frame. But this is clear, and the reason for this is that the two partial light rays meet at the same event E_2 – independent of the observer.

Why is this the case? The point is that the plane wave along the light ray has a constant phase. To demonstrate this, we take a plane wave with wavevector $\boldsymbol{k}$ that has the phase velocity $\boldsymbol{v}_{\mathrm{p}} = c\boldsymbol{e}_k$ and a light ray given by $\boldsymbol{r}(t) = \boldsymbol{v}_{\mathrm{p}}t + \boldsymbol{r}_0$, which yields $\boldsymbol{r} - \boldsymbol{e}_k ct = \boldsymbol{r}_0$. If we multiply this with $\boldsymbol{k}$, we get $\boldsymbol{kr} - ckt - \boldsymbol{kr}_0 = 0$ or $\boldsymbol{kr} - \omega t = \boldsymbol{kr}_0 \equiv \varphi_0$. Therefore, the phase of a plane wave on the light ray in spacetime is constant.

Chapter 13
Energy and Momentum

Our investigations into special relativity so far have been centered around space and time. Next, we consider **energy** and **momentum**.

We will find out that the definitions of the kinetic energy $E_{\text{kin}} = m\boldsymbol{v}^2/2$ and the momentum $\boldsymbol{p} = m\boldsymbol{v}$ from classical mechanics do not satisfy the demands of special relativity and must be modified. In special relativity, we have other formulas for the energy and the momentum, and we sometimes refer to them as the **relativistic kinetic energy** and the **relativistic momentum**. The relativistic energy is the "correct" energy and the classical expression is just an approximation of it. The same holds for the momentum. For this reason, we will identify the classical expressions with a subscript "cl" and, when referring to them, we will explicitly talk about the *classical* energy or momentum. In other words, we have $\boldsymbol{p}_{\text{cl}} = m\boldsymbol{v}$ and $E_{\text{cl,kin}} = m\boldsymbol{v}^2/2$. The names E, E_{kin}, $\boldsymbol{p}$, etc., from now on, will be associated with the relativistically correct quantities for energy and momentum. We will consequently call them the *energy* and the *momentum*, and only sporadically, for clarity, add the adjective "relativistic".

In the subsequent sections, we derive the expressions for the relativistic energy and momentum. We start in Sect. 13.1 with the (relativistic) energy and present Einstein's original derivation [Einstein05b] from the year 1905. This is a very short and ingenious derivation that, in Sect. 13.2, we complement with a discussion on what the mysterious "conversion of mass into energy" means. For the case of the (relativistic) momentum, a similar derivation was developed by Lewis and Tolman [LewisTolman09] and published in 1909. We present it in Sect. 13.3. After deriving the expression for the (relativistic) energy and the (relativistic) momentum, we illuminate their interplay in Sect. 13.4 and show that the energy and the momentum together form a four-vector. Finally, in Sect. 13.5, we illuminate the role of the conservation laws, which play a central role in physics.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_13

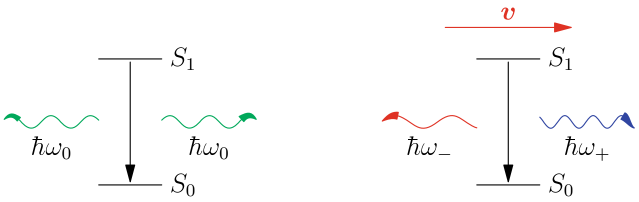


Fig. 13.1 To the equivalence of mass and energy. Left: Alice's perspective, atom at rest. Right: Bob's perspective, atom moves with velocity $\boldsymbol{v}$ to the right

13.1 The Relativistic Energy

The most popular formula of physics is probably Einstein's $E = mc^2$. We dedicate this section to it.

13.1.1 Gedanken Experiment

To introduce the topic, we follow Einstein's steps in his original publication [Einstein05b] and consider the following experiment.

When we measure the energy of an atom, we cannot get an arbitrary value, but just one out of a discrete set of values—this is directly reflected in the lines of an emission spectrum of a gas. In this vein, Einstein, in his derivation, considered an idealized atom that can only be in one of two possible *(internal) states*, the *ground state* S_0 (the state with the lowest energy) or the *excited state* S_1. He further supposed that, when the atom is in the excited state, it can "jump" into the ground state while simultaneously emitting two photons (light particles) of the same frequency in exactly opposite directions (see Fig. 13.1, left). For this reason, when the atom is at rest before the emission, it is also at rest after the emission: when emitted, the photons push the atom with the same force in opposite directions, and therefore the total force on the atom vanishes.[1]

We proceed by discussing the emission event from two different points of view: once in Alice's inertial frame, where the atom is at rest, and once in Bob's inertial frame, where the atom moves with velocity $v > 0$ in the direction of the emission of the right photon (therefore, Bob moves with velocity $-v$ relative to Alice). The central finding will be that, for Bob, *the photons take more energy away from the atom* than for Alice. In both inertial frames, however, the total *energy is conserved.* Therefore, from Bob's point of view, the atom looses more energy in the emission event than for Alice. We will see that the classical kinetic energy $E_{\text{cl,kin}} = mv^2/2$ is not consistent with this behavior.

[1] Some years later, quantum theory, would confirm Einstein's idea of how atoms and light interact.

Alice's perspective. First, we consider the emission event from Alice's point of view, i. e., in the **rest frame of the atom**. The energy of the resting atom in the ground state S_0 will be denoted by $E_0(0)$ and that in the excited state S_1 by $E_1(0)$, respectively (in parenthesis, we put the velocity of the atom, which, for Alice, vanishes). The difference is $\Delta E(0)$. The frequency of the photons is ω_0. In quantum theory, we learn that a photon of this frequency has an energy of $\hbar\omega_0$, where $\hbar$ is the Planck constant, the fundamental physical constant, which is of utmost importance in quantum theory. Before the emission event, the total energy was $E_1(0)$, and after it, the total energy is composed of the energy of the atom and the energies of the two photons and is $E_0(0) + 2\hbar\omega_0$. Therefore, we have the **energy balance**

$$\Delta E(0) = E_1(0) - E_0(0) = 2\hbar\omega_0. \tag{13.1}$$

Bob's perspective. Now, we discuss the **moving atom**. To describe the atom's *state*, we now need to specify whether it is in (internal) state S_0 or S_1 *and* its velocity. The energy of the atom will now depend on both. That's why we call the state S_i of the atom the *internal state* and use the term *state* for its internal state plus its velocity.

After this clarification, we consider the emission event from Bob's point of view. The atom before the emission event is in the excited state S_1 and moves with velocity v. We denote its energy by $E_1(v)$. Then, it emits the two photons, and subsequently enters into the ground state S_0 and has the same velocity v as before,[2] along with the energy $E_0(v)$. Let $\Delta E(v)$ again denote the difference between these energies. Because of the (longitudinal) Doppler effect (9.4), the two photons (from Bob's point of view) now have frequencies that are different from ω_0, and also differ mutually. The energy balance becomes

$$\begin{aligned} \Delta E(v) = E_1(v) - E_0(v) &= \hbar\omega_- + \hbar\omega_+ \\ &= \hbar\sqrt{\frac{1 - v/c}{1 + v/c}}\,\omega_0 + \hbar\sqrt{\frac{1 + v/c}{1 - v/c}}\,\omega_0 \\ &= 2\gamma_v\hbar\omega_0. \end{aligned} \tag{13.2}$$

As the factor γ_v is larger than 1 for $v \neq 0$, we have $\Delta E(0) < \Delta E(v)$. In other words: described in a system where the atom moves, the *photons "extract" more energy from the atom* than from the resting atom! And this in spite of the same change of the atom's internal state. We conclude that *the difference of the atom's energy in the internal states S_1 and S_0 depends on its velocity.*

Why did Einstein choose this exact Gedanken experiment to derive the relativistic formula for the energy? One ingredient of the Gedanken experiment is to describe an experiment from different inertial frames. This is nothing new here, we have done this in almost all derivations so far. The reason for this is that this allows us to implement Einstein's principle of relativity. A challenge that Einstein had in his derivation is

[2] If it changed its velocity for Bob, it could not stay at rest for Alice.

$$\begin{tikzcd}
 & E_1(0) & E_1(v) & \\
 & \bullet \arrow[r] & \bullet & \\
\Delta E(0) = 2\hbar\omega_0 & \downarrow & \downarrow & \Delta E(v) = \gamma_v \Delta E(0) = 2\gamma_v \hbar\omega_0 \\
 & \bullet \arrow[r] & \bullet & \\
 & E_0(0) & E_0(v) &
\end{tikzcd}$$

Fig. 13.2 States and transitions of the atom in Einstein's Gedanken experiment

that, at that point, he was only in possession of the relativistic kinematics: how space and time behave. But there was no obvious relation of the energy to the results up to that point. The trick that he used was to relate the energy to a kinematic quantity: the frequency of the emitted photons. The transformation behavior of frequencies was clear, and he related this to the concept of energy via the equation $E = \hbar\omega$ from quantum theory. Also essential to his Gedanken experiment was the fact that the system (atom) has a discrete energy spectrum, because energy differences are then easy to identify.

13.1.2 The Relativistic Energy

Now, we need to find the concrete expression for the energy $E_i(v)$, which must fulfill the requirements that emerged from our Gedanken experiment and which are summarized in Fig. 13.2.

Let us start by repeating the gist of the findings from Einstein's Gedanken experiment. In Fig. 13.2, the four bullets represent the states of the atom: the *atom at rest* corresponds to the two bullets on the left and the *moving atom* corresponds to the two bullets on the right. The lower two bullets correspond to the ground state S_0 and the upper two bullets to the exited state S_1. Next to each state/bullet is its energy. The downward arrows indicate a transition from the excited state to the ground state (without changing the velocity v), and next to them is the transition energy.

The energy $E_i(0)$ with $i = 0, 1$ is the **rest energy** of the atom and *depends on the internal state*. We define the **(relativistic) kinetic energy** $E_{\text{kin},i}$ of the atom in the internal state i as the difference between the energy and the rest energy:

$$E_{\text{kin},i}(v) := E_i(v) - E_i(0).$$

It *depends on both, the velocity and the internal state*.

To summarize, we have to find the function $E_i(v)$ that obeys the following conditions[3]:

[3] Note that this is a purely mathematical problem, there's no physics in it. Condition 1 is a linear system of two equations, and even if we consider $E_0(0)$ as given, it has three unknowns ($E_1(0)$, $E_0(v)$, $E_1(v)$), and therefore does not have a unique solution. Condition 2 is absolutely essential.

1. The two relations
$$\begin{aligned} E_1(0) - E_0(0) &= 2\hbar\omega_0, \\ E_1(v) - E_0(v) &= 2\gamma_v\hbar\omega_0 \end{aligned} \tag{13.3}$$
hold.
2. For small velocities, $E_{i,\mathrm{kin}}(v) = E_i(v) - E_i(0)$ becomes equal to the classical expression $mv^2/2$.

First try. To find the expression for $E_i(v)$, we can first try the obvious thing: an ansatz that is similar to that in classical physics and is composed of the rest energy $E_i(0)$ of the object that only depends on its internal state *plus* a kinetic energy $E_{\mathrm{kin}}(v)$ that *only* depends on the velocity v of the object, but not on its internal state:

$$E_i(v) = E_i(0) + E_{\mathrm{kin}}(v). \qquad \text{(this is just a try)}$$

From this follows

$$\Delta E(v) = E_1(v) - E_0(v) = E_1(0) - E_0(0) = \Delta E(0),$$

which is in contradiction to the fact in (13.3) that, at different velocities, the photons extract different amounts of energy from the atom (see Fig. 13.2).

Second try. Therefore, the dependence of the energy on the internal state S_i and the velocity v cannot be separated additively. We make the multiplicative ansatz

$$E_i(v) = E_i(0) \cdot f(v), \tag{13.4}$$

where $f(v)$ is a continuous function of v and, according to our condition 2, has to fulfill $\lim_{v\to 0} f(v) = f(0) = 1$. Note that the arrows pointing to the right in the diagram in Fig. 13.2 now indicate a multiplication with $f(v)$.

Then, as the diagram nicely shows, $E_1(v)$ can be calculated from $E_0(0)$ in two different ways, namely,

$$\begin{aligned} E_1(v) &= (E_0(0) + \Delta E(0)) \cdot f(v), \\ E_1(v) &= E_0(0) \cdot f(v) + \gamma_v \Delta E(0). \end{aligned}$$

From this, we get $f(v) = \gamma_v$, and eventually

$$E_i(v) = \gamma_v E_i(0). \tag{13.5}$$

This is the sought-for expression for the energy of the atom when it is in the internal state i and moving with velocity v. We discuss several aspects of this formula now.

Size of rest energy and zero point of the energy. In classical physics, the total energy of an object is only defined up to an additive constant; there is *no fixed zero* *point on the energy scale*. Often, the total energy of an object is divided into different types of energy. Usually, this is the *internal energy* (which, e. g., may depend on the temperature of the object), the *potential energy* (which depends on the location of the object in a force field, like the gravitational field or the electromagnetic field) and the *kinetic energy* (which depends only on the object's velocity). While the kinetic energy has a zero point (it vanishes for $v = 0$), the other energy types do not. The potential energy of an object in a gravitational field at a certain location is not defined, only the *difference* between the potential energies of an object at two different location is.

The Eq. (13.5), however, fixes the zero point for the energy scale and is not invariant upon a scale shift. If we substitute $E \to E + C$, where C is an arbitrary constant, we get $E_i(v) = \gamma_v E_i(0) + (\gamma_v - 1)C$, which has this extra term $(\gamma_v - 1)C$, and is therefore different from (13.5).

Where is this zero point located? Consider the atom in internal state S_0. For our ansatz, its typical kinetic energy is given by

$$E_{\text{kin},0}(v) = E_0(v) - E_0(0) = (\gamma_v - 1)E_0(0).$$

For small velocities, this becomes

$$E_{\text{kin},0}(v) \approx E_0(0)\frac{v^2}{2c^2}. \tag{13.6}$$

We can now compare the rest energy $E_0(0)$ with a "typical" kinetic energy and get

$$\frac{E_{\text{kin},0}(v)}{E_0(0)} \approx \frac{v^2}{2c^2}.$$

Because of $v \ll c$, $v^2/(2c^2)$ is very small, and therefore the rest energy $E_0(0)$ of an object is much larger than its kinetic energy. The kinetic energy of an object that moves with $v = 108\,\text{km/h} = 30\,\text{m/s}$ is $2c^2/v^2 = 2 \cdot 10^{14}$ times smaller than its rest energy! Its rest energy is huge!

Relation between mass and energy. So far, we have not used the *mass*, therefore, we have to introduce it somehow. But this is easy now. We already saw, in (13.6), what the expression of the relativistic kinetic energy for small velocities looks like. Clearly, this must be the same as the well-known expression $E_{\text{cl,kin}} = mv^2/2$ for the kinetic energy from classical mechanics. Comparing these two expressions yields

$$E_0(0) = mc^2,$$

which is **Einstein's famous formula**. It says that the mass of an object is nothing but its rest energy divided by the constant c^2.

Remember that, in the derivation of (13.6), we assumed the atom to be in internal state S_0. To get rid of this restriction, we now *define the* ***mass*** *as the rest energy divided by* c^2, which makes the mass of the atom depend on the internal state S_i of the atom:

$$m_i := E_i(0)/c^2. \tag{13.7}$$

The mass, by definition, does not depend on the velocity v of the atom. Moreover, the dependency on the internal state is very small and usually plays no role. This is why classical physics can live with a mass that does not depend on the internal state.

To demonstrate that the change of the mass $\Delta m = \Delta E(0)/c^2$ due to a change of the internal state of an object is much smaller than the mass $m = E(0)/c^2$ itself, we calculate the ratio

$$\frac{\Delta m}{m} = \frac{\Delta E(0)}{E(0)}$$

for the hydrogen atom.

The rest energy of the hydrogen atom is[4]

$$E(0) = m_\mathrm{p}c^2 + m_\mathrm{e}c^2 = (938.272 + 0.511)\,\mathrm{MeV} = 938.783\,\mathrm{MeV},$$

and is given here in electronvolts (eV), which is the energy unit typically used in atomic and nuclear physics.

On the other side, the maximal change of the rest energy by a change of the internal state while still not wresting away the electron from the hydrogen nucleus is the hydrogen atom's ionization energy with the value of $E_\mathrm{ioniz} = 13.6\,\mathrm{eV}$. Therefore, we get

$$\frac{\Delta E(0)}{E(0)} = \frac{E_\mathrm{ioniz}}{E(0)} = 1.45 \cdot 10^{-8}.$$

So, the ratio of a possible mass change to the mass itself is very small. The mass of an object, for all practical purposes, does not change when the object's internal state changes (e. g., through heating).

The dependence of the (relativistic) energy of an object on its mass now is easy to derive. We just have to plug the definition of mass (13.7) into (13.5) and drop the index i.

An object of mass m that moves with velocity v has the (velocity-dependent) **(relativistic) energy**

$$E = \gamma_v mc^2. \tag{13.8}$$

Here, m is the **mass** that depends on the internal state of the object but, not on its velocity.

With the **(relativistic) kinetic energy** defined by

$$E_\mathrm{kin} = (\gamma_v - 1)mc^2,$$

[4] As the rest energy depends on the internal state, we should specify the latter: the given energy corresponds to the state in which the electron is very far from the proton and at rest.

we can write $E = mc^2 + E_{\text{kin}}$.

Einstein's famous formula

$$E = mc^2$$

refers to the **rest energy** of the object, and the mass m depends on the internal state of the object.

Keep in mind that, in (13.8), $v < c$ must hold. For objects that move with the speed of light, this equation is not valid!

13.1.3 Again: Speed of Light as Maximum Velocity

Remember again the plot of the factor γ_v versus the velocity v (see Fig. 8.3). The the larger the factor becomes, the closer v approaches c and eventually goes to infinity. This is also valid for the energy of a moving object. The closer its velocity gets to c, the faster its energy increases. *The speed of light itself cannot be reached by an object with non-vanishing mass*, otherwise it would have an infinitely large energy.

This brings us back to the Bertozzi experiment (Sect. 2.1), and now we are able to derive the relativistic formula given in Fig. 2.3 and Exercise 2. We start with $E = \gamma_v mc^2$. Reordering and squaring yields $\gamma_v^{-2} = (mc^2/E)^2$ or $(v/c)^2 = 1 - (mc^2/E)^2$. Now, we put $E = mc^2 + E_{\text{kin}}$ for the energy and substitute the mass with the electron mass m_{e}, eventually arriving at the formula.

In Sect. 13.4, we will show that *massive* particles (with $m > 0$) can come arbitrarily close to the speed of light, but never can reach it, so $v < c$ in this case. *Massless* particles (with $m = 0$), however, can only exist with $v = c$. The most important massless particle is the photon.

Exercise 54: In the final configuration level of the *Large Hadron Collider (LHC)* at CERN in Geneva, protons are accelerated to a velocity of 99.9999991% of the speed of light. Calculate the ratio between their energy and their rest energy.

Exercise 55: Show the rightmost equation sign in (13.2).

Exercise 56: Consider an ideal gas consisting of $N_{\text{A}} \approx 6 \cdot 10^{23}$ point-like particles with the mass m. As James Clerk Maxwell discovered (you know him from the Maxwell equations), at the temperature T, the particles have an average kinetic energy of $E_{\text{kin}} = \frac{1}{2}m\langle \boldsymbol{v}^2\rangle = \frac{3}{2}k_{\text{B}}T$. Here, $\langle \boldsymbol{v}^2\rangle$ is the average of the squares of the individual velocities and k_{B} is the *Boltzmann constant*, a fundamental physical constant named after Ludwig Boltzmann. Determine the dependency of the mass of the gas on the temperature.

13.1.4 The Discussion About the "Relativistic Mass"

In his original publication, Einstein writes (see [Einstein05b], the notation was adapted): *„Die Masse eines Körpers ist ein Mass für dessen Energieinhalt; ändert sich die Energie um ΔE, so ändert sich die Masse in demselben Sinne um $\Delta E/c^2$ …“*.[5] Thus, Einstein understood the mass of an object to be a variable quantity that depends on the energy and therefore changes its value with the velocity (and not only with the internal state). His relativistic mass m_{rel} is what, in this book, we denote by E/c^2 or $\gamma_v m$. Together with the relativistic mass, one also uses the concept of *rest mass* m_0, which corresponds to our concept of an (invariant) mass m (which needs no subscript).[6]

There are several factors that speak against using the (velocity-dependent) relativistic mass. First, because of $E = m_{\text{rel}}c^2 = \gamma_v mc^2$, the relativistic mass would be nothing but the energy of the object, and one might wonder whether two quantities are needed for one and the same property of an object. Occam's razor suggests disposing of one of the concepts. A second reason is that, in the mathematical formulation of special relativity, the language of tensors is used (this is an advanced concept that we do not use). In this language, there are only a few classes of different quantities and they are defined with respect to their transformation behavior under a Lorentz transformation. These classes are *scalars* (which are invariant), *four-vectors* like $(ct, \boldsymbol{x})$, $(\omega/c, \boldsymbol{k})$, $(E/c, \boldsymbol{p})$, *second-rank tensors*, etc. A variable relativistic mass has no place here. If one abandons the relativistic (velocity-dependent) mass, that only leaves the rest mass m_0, which, without danger of confusion, can be called the mass m. One could argue that the particular mathematical formalism (tensors) used to describe physical phenomena must not have any implication for the physical concepts. This is true, but the tensors originate from a very physical concept, which is the space and time symmetries.

In later years, Einstein himself rejected the concept of the relativistic (velocity-dependent) mass. On June 19, 1948, he writes in a letter to Lincoln Barnett (his mass M corresponds to m_{rel})[7]: *"It is not good to introduce the concept of the* mass $M = m/(1 - v^2/c^2)^{1/2}$ *of a moving body for which no clear definition can be given. It is better to introduce no other mass concept than the 'rest mass' m. Instead of introducing M it is better to mention the expression for the momentum and the energy of a body in motion."*

[5] "The mass of an object is a measure for its energy content; if the energy changes by ΔE, then the mass changes in the same sense by $\Delta E/c^2$…".

[6] The (invariant) mass of an object, by definition, does not depend on the velocity of the object. It does, however, depend on the internal energy of the object. If we heat the object, its (invariant) mass increases.

[7] Cited in [Okun1989]. The author L. B. Okun notes: "Einstein wrote in German; the letter was typed and sent in English.".

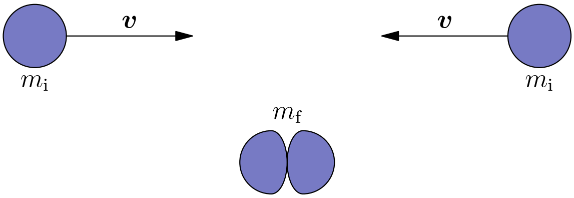


Fig. 13.3 Collision experiment to demonstrate the mass defect

13.2 "Conversion" of Mass into Energy: Mass Defect

In special relativity, the energy of an object in an inertial frame via $E = \gamma_v mc^2$ depends on its mass m and its velocity v. *A consequence of this is that, in a collision, the total energy is conserved but the sum of the masses no longer is*. This can be seen in a very nice way in the **perfectly inelastic collision**. Two equal particles with the same initial mass m_i and the same magnitude of velocity move on one (straight) line toward each other and collide[8] (see Fig. 13.3).

The energy before the collision is

$$E_\mathrm{i} = E_{\mathrm{i},1} + E_{\mathrm{i},2} = 2\gamma_v m_\mathrm{i} c^2.$$

After the collision, the resulting merged object is at rest and has the mass m_f, and therefore the energy

$$E_\mathrm{f} = m_\mathrm{f} c^2.$$

From the energy conservation $E_\mathrm{i} = E_\mathrm{f}$, it follows that

$$m_\mathrm{f} = 2\gamma_v m_\mathrm{i} > 2m_\mathrm{i}.$$

Therefore, the mass of the merged object after the collision is *larger* than the sum of the masses of the individual particles before the collision. **The kinetic energy has been converted completely into mass**. In the logic of Sect. 13.1, the kinetic energy has been converted into excitations of the merged object (it became hotter, etc.), and therefore its mass is now larger.

Another example is the **energy release in the Sun**, where, according to the reaction equation

$$4\,{}^1_1\mathrm{H} \longrightarrow {}^4_2\mathrm{He} + 2e^+ + 2\nu_e + 2\gamma \qquad (13.9)$$

via several intermediate steps, four hydrogen nuclei (${}^1_1$H) are converted into one helium nucleus (${}^4_2$He), two positrons (e^+), two neutrinos (ν_e) and light.[9] The mass[10]

[8] The indices "i" and "f" stand for "initial" and "final" and refer to the situation before and after the collision, respectively.

[9] In particle physics, photons are denoted by γ. This has nothing to do with our factor γ_v.

[10] $1\,\mathrm{u} = 1.6605 \cdot 10^{-27}$ kg is the *atomic mass unit*. You will certainly have already learned this in an atomic physics course or in chemistry.

of a hydrogen nucleus is $m({}^1_1\mathrm{H}) = 1.0078\,\mathrm{u}$, that of a helium atom is $m({}^4_2\mathrm{He}) = 4.00260\,\mathrm{u}$ and the positron mass is $m(e^+) = 0.0005\,\mathrm{u}$. The mass of the neutrinos is insignificant. Therefore, the mass balance is

$$\begin{aligned} m_\mathrm{i} &= 4\,m({}^1_1\mathrm{H}) = 4.0312\,\mathrm{u}, \\ m_\mathrm{f} &= m({}^4_2\mathrm{He}) + 2m(e^+) = 4.0036\,\mathrm{u}. \end{aligned}$$

Hence, in the reaction (13.9), the mass difference of $\Delta m = 0.0276\,\mathrm{u}$ is converted into (kinetic) energy (thus, heat) and radiation.

If one mole of hydrogen (with about a gram of mass) is converted into helium, an amount of

$$\Delta E = N_\mathrm{A} \cdot \Delta m c^2 = 0.0276\,\mathrm{g} \cdot c^2 \approx 2.5\,\mathrm{TJ}$$

of energy in the form of heat and radiation is released (N_A is the Avogadro constant, which is $N_\mathrm{A} \cdot 1\,\mathrm{u} = 1\,\mathrm{g}$).

Exercise 57: Germany has a primary energy consumption of about 15 PJ ($1\,\mathrm{PJ} = 10^{15}\,\mathrm{J}$), with the private sector's share being about 25%. How much is the primary energy consumption of the private sector per capita (Germany has about 83 million residents)? How many residents could be provided with energy for one year from the nuclear fusion of 1 mol of hydrogen? For how long could the whole primary energy consumption of Germany be covered? (The numbers are approximate and for the year 2000.)

Exercise 58: To release 2.5 TJ by burning coal, one needs about 85 tons of it. Show that this is true. (According to Wikipedia,[11] by burning 1 ton of coal, an energy of 29.3 GJ is released.)

13.3 The Relativistic Momentum

Now, we derive an expression for the (relativistic) momentum. Consider the collision of two particles 1 and 2 (see Fig. 13.4, left side).[12] The particles are "equal", in particular, they have the same mass. We assume that they move toward each other and have the same magnitude of velocity. Therefore, the total momentum vanishes. The inertial frame in which it vanishes is called the **center-of-mass frame** S. After the collision, the particles may travel in different directions, but these must be opposite directions and the velocities of the particles must have the same magnitude again. Otherwise, the total momentum after the collision would not vanish and the

[12] In the literature, one finds different derivations of the relativistic momentum. The presumably clearest one from a logical point of view, which, at the same time, is also the simplest, is that presented here and also used by Richard P. Feynman in his famous Lectures on Physics. It originates from Gilbert N. Lewis and Richard C. Tolman [LewisTolman09].

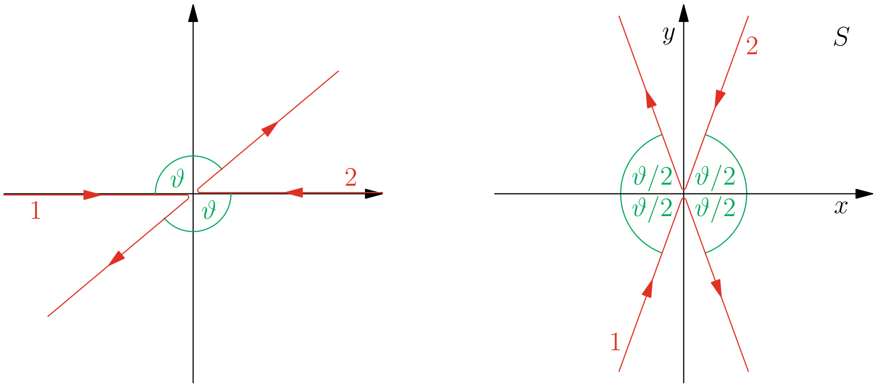


Fig. 13.4 Collision of two equal particles in the center-of-mass frame (with the same magnitude of the velocity). Right: The same as on the left, but with rotated coordinate systems

momentum conservation would be violated. We denote the scattering angle by ϑ (see Fig. 13.4, left side).

To ease our task, we rotate the coordinate system in such a way that the trajectories of each particle before and after the collision result from a reflection at the y-axis and the trajectories of particles 1 and 2 (but inverted) result from a reflection at the x-axis. In this coordinate system, the x-component of the momentum of particle 1 before and after the collision is the same (this also holds for particle 2). The y-component of the momentum, however, changes its sign at the collision (the same also holds here for particle 2).

The relation between the (relativistic) momentum and the velocity is still unknown to us. In classical mechanics, $\boldsymbol{p}_{\text{cl}} = m\boldsymbol{v}$. For the **relativistic momentum**, we make a similar ansatz,

$$\boldsymbol{p} = f(v)m\boldsymbol{v}, \tag{13.10}$$

as for the relativistic energy (for symmetry reasons, the relativistic momentum must be parallel to the velocity). Here, v is the magnitude of the velocity. Because the relativistic momentum for small velocities must agree with the classical momentum, $f(v) \to 1$ for $v/c \to 0$ is necessary.

In the inertial frame S, the center-of-mass frame, that we have used so far, it is obvious that, due to the existing symmetries, the momentum is conserved. Now, we go to a different inertial frame and describe the collision from this new point of view (in the derivation of the relativistic energy in Sect. 13.1, we did exactly the same thing). Momentum conservation must also hold in the new inertial frame, otherwise, the relativity principle would be violated. This requirement will lead us to $f(v)$.

Let the inertial frame S' with its coordinate axes x' and y' be such that the x- and the x'-axis coincide, the y- and the y'-axis are parallel and the x-component of the velocity of particle 1 vanishes. In this inertial frame, the collision appears as shown in Fig. 13.5 on the left side. Let the velocity of particle 1 in S' be w. Furthermore,

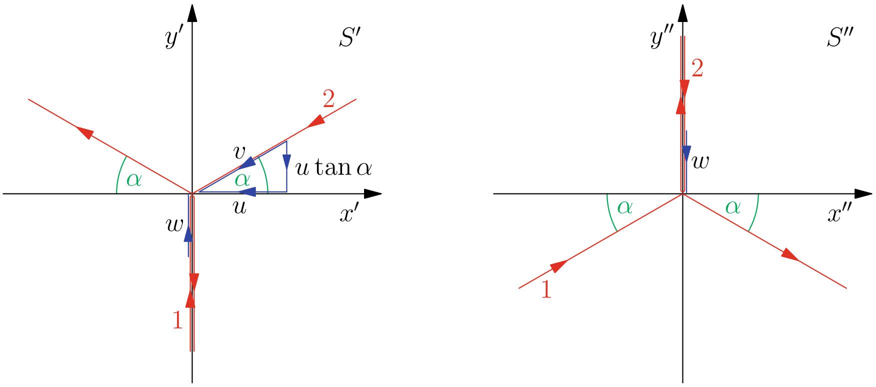


Fig. 13.5 The same collision as in Fig. 13.4. Left: In the inertial frame S', where the x-component of the velocity of particle 1 vanishes. Right: In the inertial frame S'', where the x-component of the velocity of particle 2 vanishes

we denote the *magnitude* of the velocity of particle 2 in S' by v, its x'-component by u and the angle between its direction of motion and the x'-axis by α. Then, the y'-component of the velocity of particle 2 is given by $u \tan\alpha$.[13] The velocity components u, v and w are defined such that they are positive.

Now, we apply **momentum conservation** for the y'-component of the momentum. The change of the momentum of particle 1, according to our ansatz (13.10), is given only by $\Delta p_{1,y'} = 2f(w)mw$ and that of particle 2 by $\Delta p_{2,y'} = 2f(v)mu\tan\alpha$ (note that according to our ansatz, the velocity *magnitude* appears in the argument of function f, and not the velocity component $u\tan\alpha$). Momentum conservation then requires that $\Delta p_{1,y'} = \Delta p_{2,y'}$, or

$$f(w)w = f(v)u\tan\alpha. \tag{13.11}$$

To determine $u\tan\alpha$, we change to another inertial frame S'', which moves with the velocity u relative to S' (again, the x'- and the x''-axis coincide and the y'- and the y''-axis are parallel) (see Fig. 13.5, right side). Therefore, the x''-component of the particle 2 vanishes. Because both particles before the collision had the same velocity, the y''-component of the velocity of particle 2 must be w.

Next comes the key step in the derivation. We have to relate $u\tan\alpha$ and w. In classical physics, these would be equal, but in relativistic physics, they are not, and the reason is time dilation.

We need to consider only particle 2 to find this relation. In S'', the y''-component of particle 2 is w. What is the y'-component of its velocity in S'? The lengths perpendicular to the relative motion of S' and S'' are the same. But for an observer resting in S', the time of particle 2 passes more slowly by the factor γ_v than its own time.

[13] In the case of small velocities, $\boldsymbol{p} = m\boldsymbol{v}$ and, because of the conservation of the y'-components of the momentum, we had $w = u\tan\alpha$. This is no longer correct for large velocities.

In total, the velocity of the particle in the y'-direction and measured in S' must be smaller by a factor of γ_u^{-1} than in S''. For that reason,

$$u \tan \alpha = \gamma_u^{-1} w.$$

Eq. (13.11) then becomes $f(w)w = f(v)\gamma_u^{-1} w$ or

$$f(v) = \gamma_u f(w).$$

From this relation, we determine the function $f(v)$, and, to achieve this, we must get rid of one occurrence of the function. We do this by letting its argument go to 0, after which the function's value goes to 1. But in doing so, we have to take into account that the velocities u, w, v are mutually dependent. Putting $v \to 0$ is not sensible, because w also then goes to 0 and f vanishes completely in the equation above. But we can put $w \to 0$ without having v disappear. The only thing we have to do is to fix u. For u fixed and $w \to 0$, because of $u \tan \alpha = \gamma_u^{-1} w$, we have $u \tan \alpha \to 0$ and, finally, $\alpha \to 0$. But if α goes to zero, v goes to u, and therefore $f(v)$ goes to $f(u)$ and we have the solution: $f(u) = \gamma_u$. Together with (13.10), we arrive at:

The **(relativistic) momentum** of an object with mass m, which moves with velocity $\boldsymbol{v}$ relative to an inertial observer, is given by

$$\boldsymbol{p} = \gamma_v m \boldsymbol{v}. \tag{13.12}$$

For small velocities $v \ll c$, we have $\gamma_v \approx 1$ and (13.12) becomes the momentum $\boldsymbol{p}_{\text{cl}} = m\boldsymbol{v}$ of classical mechanics.

Note that, in (13.12), $v < c$ must hold. For objects that move with the speed of light, the equation is not applicable anymore.

We discuss again the logic of the derivation. If, in a physical process described from a particular inertial frame, the momentum is conserved, then it must be conserved in all inertial frames, otherwise, we would be able to distinguish the inertial frames, in contradiction to the principle of relativity. In the case of a collision of two particles, the total momentum is simply given by the sum of the momenta of the particles. Thus, the expression for the momentum must be such that, when it is transformed to another inertial frame, it must be conserved there as well. It must be form-invariant under coordinate transformations. In classical mechanics, one uses the Galilei transformation and, with respect to it, the expression $\boldsymbol{p} = m\boldsymbol{v}$ is indeed form-invariant. In special relativity, however, we must use the Lorentz transformation. Then, the expression $\boldsymbol{p} = m\boldsymbol{v}$ is not form-invariant anymore. We must find another expression. It is of great help that this expression for the relativistic momentum for small velocities must become equal to the classical momentum. With the construction above, we have seen that (13.12) is the correct generalization of the classical momentum to special relativity.

Note that this has nothing to do with the question as to whether the total momentum is conserved or not. If the *classical* momentum were conserved for large velocities, we would indeed have a problem. In that case, the relativistic momentum would not be conserved and either the principle of momentum conservation or special relativity would be wrong. Fortunately, the relativistic momentum is conserved. This is demonstrated day by day in millions of particle collisions in particle accelerators.

We will come back to this in Sect. 13.5.

13.4 Interplay of Energy and Momentum

Suppose an object moves with velocity u relative to the inertial observer Alice. Thus, (for Alice), it has the energy $E = \gamma_u mc^2$ and the momentum $p = \gamma_u mu$. Bob moves with velocity v relative to Alice. Thus, the object moves relative to Bob with the velocity $u' = u \oplus (-v)$ and (again for Bob) has the energy $E' = \gamma_{u'} mc^2$ and the momentum $p' = \gamma_{u'} mu'$. How can the energy and momentum of an object in one inertial frame be directly transformed to the related quantities in another inertial frame? In other words: what are the Lorentz transformations for energy and momentum?

This question has a surprisingly simple answer. We need the γ-formulas (12.5) (which, as you remember, are nothing but the Lorentz transformation of the velocity). Multiplying the first of these formulas on both sides with m, one gets $p' = \gamma_v \cdot (\gamma_u mu - v\gamma_u m)$. The first term in parenthesis is equal to p and the second one to vE/c^2. In total, we have $p' = \gamma_v \cdot (p - vE/c^2)$. Now, we multiply the second of the formulas above with mc^2 and get $E' = \gamma_v \cdot (E - vp)$. We arrive at:

The **Lorentz transformation** for the **energy** and the **momentum** is given by

$$\begin{aligned} p' &= \gamma_v \cdot (p - \frac{v}{c^2} E), \\ E' &= \gamma_v \cdot (E - vp). \end{aligned} \tag{13.13}$$

These are the sought-for transformation formulas for the (relativistic) energy and momentum. If you compare them to the Lorentz transformation (11.9), you see that, indeed, in the Lorentz transformation, one simply has to perform the replacements $t \to E/c^2$ and $x \to p$ to get the transformation for energy and momentum. This means that $(E/c, \boldsymbol{p})$ is a four-vector, and we have

$$\begin{pmatrix} E/c \\ \boldsymbol{p} \end{pmatrix} = m \begin{pmatrix} c \\ \boldsymbol{u} \end{pmatrix}.$$

The **energy-momentum four-vector is just the mass m times the velocity four-vector**.

Remember that the invariant $c^2t^2 - x^2$ follows directly from the Lorentz transformation. In the same way, we see that the expression $(E/c)^2 - p^2$ is invariant. In the rest frame of an object, $E = mc^2$ and $p = 0$, so the invariant has the value $(mc^2)^2$. From that follows:

Energy-momentum relation: The relativistic energy E of an object of mass m is related to the relativistic momentum $\boldsymbol{p}$ via

$$E^2 = m^2c^4 + \boldsymbol{p}^2c^2. \tag{13.14}$$

This expression, in comparison to our previous formulas for the relativistic energy (13.8) and the relativistic momentum (13.12), has the big advantage that it is valid also for $v = c$ and even for objects with vanishing mass, i. e., for the case $m = 0$. Then, $E = |\boldsymbol{p}|c$. In this case, objects necessarily move with the speed of light.[14] Objects with $m = 0$ and $v < c$ are forbidden in special relativity. Photons are (quantum) objects with $m = 0$ and, therefore, $v = c$.

Then, we see that there are two types of object:

(Rest) mass	Velocity	Energy-momentum relation	Term
$m > 0$	$v < c$	$E(\boldsymbol{p}) = \sqrt{m^2c^4 + \boldsymbol{p}^2c^2}$	massive
$m = 0$	$v = c$	$E(\boldsymbol{p}) = \lvert\boldsymbol{p}\rvert c$	massless

Suppose that $m > 0$. Then, $E(\boldsymbol{p}) = \sqrt{m^2c^4 + \boldsymbol{p}^2c^2}$ (see Fig. 13.6). For $\boldsymbol{p} = 0$, thise becomes equal to the rest energy $E(0)$ and, in the case of very large momenta, (13.14) becomes $E = |\boldsymbol{p}|c$. The energy then grows linearly with the momentum.

Based on our findings, we can adapt Minkowski's words from the end of Sect. 11.3 and say:

> The views of *energy* and *momentum* which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth, *energy* by itself, and *momentum* by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.

[14] To show this, we need a concept that you have yet to learn: one gets the velocity of an object by deriving the energy-momentum relation $E(\boldsymbol{p})$ with respect to the momentum. We restrict our discussion to one dimension. In classical mechanics, the (kinetic) energy is $(m/2)v^2 = p^2/(2m)$ or $E_{\text{cl,kin}}(p) = p^2/(2m)$. Calculating the derivative of this gives us $dE(p)/dp = p/m$, which is the velocity. If we, however, differentiate the relativistic relation, we get $dE/dp = pc^2/\sqrt{m^2c^4 + p^2c^2}$, which, for $m = 0$, becomes $dE/dp = \pm c$. Therefore, massless objects move with the speed of light.

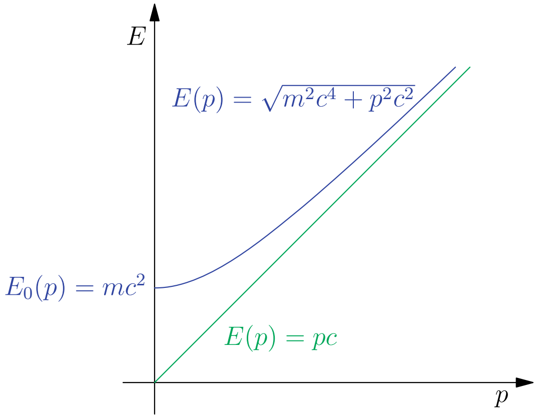


Fig. 13.6 Energy-momentum relation for an object with $m \neq 0$ ("massive particle") (blue) and a massless object (green)

Exercise 59: From Einstein's arguments, we learned that the energy as a function of the velocity is given by

$$E(v) = E(0) + (\gamma_v - 1)mc^2, \tag{13.15}$$

where the rest energy $E(0)$ could be an arbitrary constant. Show that $E(0)$ must be equal to mc^2 if the expression (13.15) has to fulfill the following transformation (see (13.13)):

$$\begin{aligned} p' &= \gamma_v \cdot \left(p - \frac{v}{c^2}E\right), \\ E' &= \gamma_v \cdot (E - vp)\,. \end{aligned} \tag{13.16}$$

You can proceed as follows: consider an object that is at rest for Alice. What is its energy according to (13.15)? What is its momentum (we don't have to know the expression for the (relativistic) momentum to answer this question)? By using (13.16), transform the quantities to Bob's inertial frame, which moves with velocity v relative to Alice. E' is the energy of the object for Bob. On the other hand, instead of making the transformation (13.16), Bob can simply use the expression (13.15) to determine the energy. What follows from that?

13.5 Energy and Momentum Conservation Laws

Einstein, in his original derivation of the expression for the (relativistic) energy (see Sect. 13.1), and Lewis and Tolman, in their derivation of the expression for the

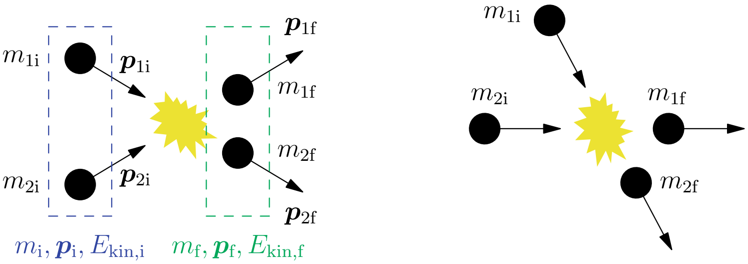

$m_{1\mathrm{i}}$
$\boldsymbol{p}_{1\mathrm{i}}$
$m_{2\mathrm{i}}$
$\boldsymbol{p}_{2\mathrm{i}}$
$\boldsymbol{p}_{1\mathrm{f}}$
$m_{1\mathrm{f}}$
$m_{2\mathrm{f}}$
$\boldsymbol{p}_{2\mathrm{f}}$
$m_\mathrm{i}, \boldsymbol{p}_\mathrm{i}, E_\mathrm{kin,i}$
$m_\mathrm{f}, \boldsymbol{p}_\mathrm{f}, E_\mathrm{kin,f}$
$m_{1\mathrm{i}}$
$m_{2\mathrm{i}}$
$m_{1\mathrm{f}}$
$m_{2\mathrm{f}}$

Fig. 13.7 Scheme of a collision of two particles. Right and left side represent the situation as seen from different inertial frames

(relativistic) momentum (see Sect. 13.3), achieved their goals in a similar way: they have constructed simple Gedanken experiments and used three ingredients[15]:

1. the principles that **energy and momentum** are **conserved**,
2. the **principle of relativity** and the consequence that, *if* energy and momentum are conserved in one inertial frame, they must also be conserved in any other inertial frame, and
3. the requirement that the expression for the (relativistic) momentum and energy must become equal to the expressions for the classical momentum and energy, respectively, in the **limiting case of small velocities**.

In this section, we show again how these ingredients work together and illuminate the prominent role of the conservation laws.

The system that we are considering is a system of **two point-like particles** upon which no forces from the outside act and that **collide** at some point (see Fig. 13.7). The system is the same as that considered in the derivation by Lewis and Tolman, but here, we allow for different masses. We will describe this experiment from Alice's point of view (left side of the figure) and from Bob's point of view (right side). Momentum, (kinetic) energy, and mass are labeled with the number of the particle to which they refer and whether they apply to before the collision (index "i" for "initial") or after it (index "f" for "final").

13.5.1 Conservation Laws

The conservation laws are included in the basic equations of a theory. Let us go back to classical mechanics and show that, indeed, the conservation laws for the classical energy and momentum follow from Newton's laws.

[15] Strictly speaking, the first ingredient is not really needed. But we have to assume that, if the conservation law is fulfilled in one inertial frame, it also will have to be fulfilled in any other inertial frame.

For the **derivation** of the conservation laws for momentum and energy, we assume that the two particles, long before they collide, do not interact at all. Then, there will be a time period when they interact. Only in this time period do forces act on these particles. After this time period, the particles do not interact anymore.

Momentum conservation. This is pretty straightforward. If, in an inertial frame, the two particles 1 and 2 interact, then, according to Newton's third law ("action equals reaction"), the force on particle 2 has the same magnitude as the force on particle 1 but points exactly in the opposite direction: $\boldsymbol{F}_1 = -\boldsymbol{F}_2$. We now consider a very short time interval Δt in which the force can be assumed to be constant. Then, the changes of the momenta due to the forces are

$$\Delta \boldsymbol{p}_1 = \boldsymbol{F}_1 \Delta t \quad \text{and} \quad \Delta \boldsymbol{p}_2 = \boldsymbol{F}_2 \Delta t = -\Delta \boldsymbol{p}_1 \ .$$

Adding both equations yields $\Delta(\boldsymbol{p}_1 + \boldsymbol{p}_2) = 0$, the momentum conservation for the considered short time interval Δt. As the whole collision process can be partitioned into small time intervals, momentum conservation has been proved for the considered case of two-particle collisions.

Energy conservation. This is a bit more tricky.[16]

Suppose a particle moves on its trajectory $\boldsymbol{r}(t)$ and a force $\boldsymbol{F}(t)$ acts on it (see Fig. 13.8, left). Then, in the short time interval from t_1 to t_2, the particle moves along the short trajectory element from $\boldsymbol{r}_1 = \boldsymbol{r}(t_1)$ to $\boldsymbol{r}_2 = \boldsymbol{r}(t_2)$. Suppose that the force $\boldsymbol{F}(t)$ is close to constant on this small trajectory element and let $\Delta \boldsymbol{r} = \boldsymbol{r}_2 - \boldsymbol{r}_1$ and $\Delta t = t_2 - t_1$. Then, the force $\boldsymbol{F}$ performs the work $\Delta W = \boldsymbol{F} \cdot \Delta \boldsymbol{r}$ on the particle (the scalar product says that only the force component parallel to the trajectory performs work, while the component perpendicular to it plays no role).

Now, we can write $\Delta W = \boldsymbol{F} \cdot (\Delta \boldsymbol{r}/\Delta t) \cdot \Delta t$. For small Δt, the fraction becomes the particle's velocity, so $\Delta W = \boldsymbol{F} \cdot \boldsymbol{v} \cdot \Delta t$. Using Newton's second law $\boldsymbol{F} = m\, d\boldsymbol{v}/dt$, we then have $\Delta W = m\dot{\boldsymbol{v}}\boldsymbol{v}\Delta t$, and this can be written as

$$\Delta W = \frac{d}{dt}\left(\frac{m}{2}\boldsymbol{v}(t)^2\right)\Delta t = \frac{dE_{\text{kin}}}{dt}\Delta t \approx \Delta E_{\text{kin}}$$

(for small time intervals Δt). Therefore, the work ΔW performed by the force $\boldsymbol{F}$ on the particle increases the kinetic energy by exactly the same amount.

Next, we consider a trajectory $\boldsymbol{r}(t)$ from t_i to t_f and $\boldsymbol{r}_\text{i} = \boldsymbol{r}(t_\text{i})$ and $\boldsymbol{r}_\text{f} = \boldsymbol{r}(t_\text{f})$, which is not necessarily small (the indices i and f refer to initial and final, respectively). Then, the change of the kinetic energy of the particle due to the work performed by the force $\boldsymbol{F}(t)$, while it moved on the trajectory, is given by

$$\Delta E_{\text{kin}} = E_{\text{kin},2} - E_{\text{kin},1} = \int \boldsymbol{F}(\boldsymbol{r})\, d\boldsymbol{r},$$

[16] The reason for this is that we want to avoid introducing the potential energy.

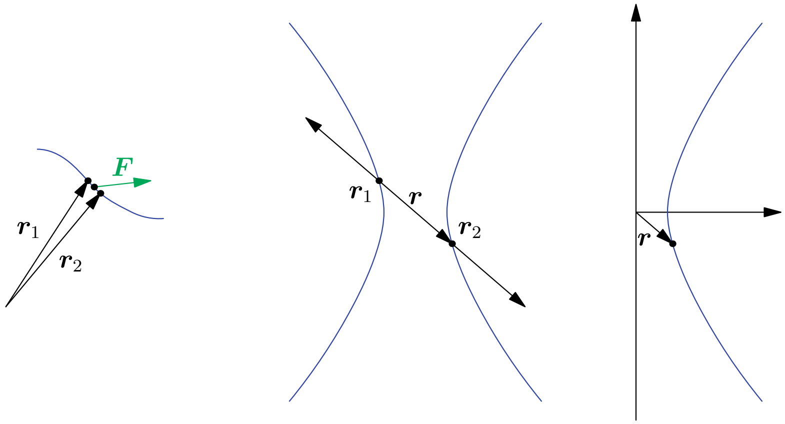


Fig. 13.8 Left: Particle moves on trajectory $\boldsymbol{r}(t)$ and force $\boldsymbol{F}(t)$ acts on it. Middle: Two partices collide. Right: Relative coordinate (scaled down by a factor of 2)

where the integral is a *curve integral* that sums up all the work ΔW performed on the particle during its whole movement from $\boldsymbol{r}_\mathrm{i}$ to $\boldsymbol{r}_\mathrm{f}$.

Now suppose that two particles 1 and 2 are colliding and $\boldsymbol{F}_i(\boldsymbol{r})$ $(i = 1, 2)$ is the force on particle i (see Fig. 13.8, middle). If we add the work performed on particle 1 and that on particle 2, we get

$$\Delta E_\text{kin,1} + \Delta E_\text{kin,2} = \int \boldsymbol{F}_1(\boldsymbol{r}_1)\, d\boldsymbol{r}_1 + \int \boldsymbol{F}_2(\boldsymbol{r}_2)\, d\boldsymbol{r}_2.$$

For the forces $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$, Newton's third law holds. The force $\boldsymbol{F}_1$ exerted by particle 2 on particle 1 is equal in magitude and points in the opposite direction than the force $\boldsymbol{F}_2$ exerted by particle 1 on particle 2: $\boldsymbol{F}_1 = -\boldsymbol{F}_2$. Furthermore, we use the relative coordinate $\boldsymbol{r} = \boldsymbol{r}_1 - \boldsymbol{r}_2$, which points from particle 2 to particle 1, and $\boldsymbol{F} := \boldsymbol{F}_1$. Thus,

$$\Delta E_\text{kin,1} + \Delta E_\text{kin,2} = \int \boldsymbol{F}\, d\boldsymbol{r}.$$

This is again a curve integral and is taken along the *relative path* $\boldsymbol{r}(t)$ (see Fig. 13.8, right). Now, the important point is that, for a large class of force fields, called *conservative forces*, this integral vanishes. We can take the integral from some time t_i way before the interaction starts until some time t_f way after it ends and then extend the integral over the curve to an integral over a closed curve in a region of space where the particles do not interact. The definition of conservative forces is that all the curve integrals of $\boldsymbol{F}\, d\boldsymbol{r}$ along closed curves vanish.

Therefore, for *conservative force fields*, we have

$$E_\text{kin,1,i} + E_\text{kin,2,i} = E_\text{kin,1,i} + E_\text{kin,2,f},$$

so the **total kinetic energy**

$$E_{\text{kin,total}} = E_{\text{kin},1} + E_{\text{kin},2}$$

is conserved in this case: the kinetic energy way before and way after the collision is the same. Collisions during which the kinetic energy is conserved are also called **elastic** collisions.

We see that momentum conservation is very general and corresponds directly to Newton's third law. The conservation of the kinetic energy in a collision of the considered type, however, requires that the force field be conservative. The force fields of the gravity force and electric forces are conservative. Therefore, if the forces $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ are gravity or electric forces, the total kinetic energy way before and way after the collision are the same. There are, however, a lot of situations when the force field is not conservative and, therefore, the kinetic energy is not conserved. Examples are:

- if there is *dissipation* in the collision, i. e., when the particles heat up or even stick together (as in Fig. 13.3), or
- if some *other motions are caused* by the collision, like a rotation or an oscillation of a particle, or
- if the particle is charged and *magnetic forces* act on it.

Note also that, *during the collision*, the kinetic energy is not conserved. If two particles of the same mass approach each other on the same line and with the same velocity magnitude, at some point, they will come to rest before they start moving away from each other. In this moment, the kinetic energy is zero. Total energy, however, is always conserved. In the case at hand, the kinetic energy of the particles far from the collision during the collision is converted into *potential energy* or *deformation energy* of the particles before it is converted into kinetic energy again.

13.5.2 Principle of Relativity

If a conservation law were to hold in one inertial frame and not hold in another one, the principle of relativity would be violated. Therefore, a conservation law must either hold in all inertial frames or in none.

Alice. We first consider the momentum. Before the collision, the momentum of particle 1 is given by $\boldsymbol{p}_{1\text{i}}$ and the momentum of particle 2 is given by $\boldsymbol{p}_{2\text{i}}$. The total momentum before the collision is then $\boldsymbol{p}_\text{i} = \boldsymbol{p}_{1\text{i}} + \boldsymbol{p}_{2\text{i}}$. After the collision, the particles have the momenta $\boldsymbol{p}_{1\text{f}}$ and $\boldsymbol{p}_{2\text{f}}$, respectively, and the total momentum is $\boldsymbol{p}_\text{f} = \boldsymbol{p}_{1\text{f}} + \boldsymbol{p}_{2\text{f}}$. The **conservation of the total momentum** means that

$$\boldsymbol{p}_\text{i} = \boldsymbol{p}_\text{f} \quad \text{or} \quad \boldsymbol{p}_{1\text{i}} + \boldsymbol{p}_{2\text{i}} = \boldsymbol{p}_{1\text{f}} + \boldsymbol{p}_{2\text{f}} \,. \tag{13.17}$$

The analogous statement holds for the energy. The **conservation of the total energy** means that

$$E_\mathrm{i} = E_\mathrm{f} \quad \text{or} \quad E_{1\mathrm{i}} + E_{2\mathrm{i}} = E_{1\mathrm{f}} + E_{2\mathrm{f}}.$$

Bob. As we mentioned earlier: if a quantity is conserved in one inertial frame, then it must also be conserved in any other inertial frame (though it has a different value in a different inertial frame). Otherwise, the principle of relativity would be violated, because certain inertial frames would be special due to the fact that the conservation of total energy and momentum holds in them. If the conservation of *total* energy and momentum holds for Alice, it must also hold for another inertial observer Bob. Due to the fact that the individual particles have a different velocity for Bob than for Alice, energy and momentum are also different for both. If we denote the quantities for Bob by a prime, *the conservation of the total momentum for Bob* is

$$\boldsymbol{p}'_\mathrm{i} = \boldsymbol{p}'_\mathrm{f} \quad \text{or} \quad \boldsymbol{p}'_{1\mathrm{i}} + \boldsymbol{p}'_{2\mathrm{i}} = \boldsymbol{p}'_{1\mathrm{f}} + \boldsymbol{p}'_{2\mathrm{f}}, \tag{13.18}$$

and *the conservation of the total energy for Bob* reads as

$$E'_\mathrm{i} = E'_\mathrm{f} \quad \text{or} \quad E'_{1\mathrm{i}} + E'_{2\mathrm{i}} = E'_{1\mathrm{f}} + E'_{2\mathrm{f}}. \tag{13.19}$$

If the transformation is such that (13.17) holds exactly when (13.18) is true, then we say that (13.17) (or (13.18), respectively) is consistent with the principle of relativity. The law of conservation of momentum, as well as that for energy must therefore be consistent with the Lorentz transformation.

Therefore, we have to show now that, if momentum conservation holds for Bob, then also holds for Alice. And thus the same for the energy. To do so, we need the transformation laws for the momentum and the energy that we already determined in the last section. If we did not have these transformation laws, we could determine them from the requirement that they have to make the conservation laws invariant—exactly as Lewis and Tolman did.

13.5.3 Classical Mechanics

First, we make sure that, in classical mechanics, (13.17) is compatible with the principle of relativity. *Note: in this Sect. 13.5.3, the quantities $\boldsymbol{p}$ and E refer to the classical momentum and energy, respectively.*

Momentum. In classical mechanics, the momentum of a particle is given by $\boldsymbol{p} = m\boldsymbol{u}$, where $\boldsymbol{u}$ is the velocity of the particle. The mass m of a particle is an invariant (it has the same value for Alice and Bob). The transformation of the velocity is easy. If the particle has velocity $\boldsymbol{u}$ in Alice's inertial frame and Bob has velocity $\boldsymbol{v}$ relative to Alice, then the particle has velocity $\boldsymbol{u}' = \boldsymbol{u} - \boldsymbol{v}$ relative to Bob. For the momentum, one gets immediately

$$\boldsymbol{p}' = m\boldsymbol{u}' = m\boldsymbol{u} - m\boldsymbol{v},$$

the **transformation law for the momentum** therefore is

$$\boldsymbol{p}' = \boldsymbol{p} - m\boldsymbol{v} \,. \tag{13.20}$$

With this transformation law, (13.17) follows from (13.18), and vice versa. From (13.18), using (13.20), we get

$$(\boldsymbol{p}_{1\mathrm{i}} - m_{1\mathrm{i}}\boldsymbol{v}) + (\boldsymbol{p}_{2\mathrm{i}} - m_{2\mathrm{i}}\boldsymbol{v}) = (\boldsymbol{p}_{1\mathrm{f}} - m_{1\mathrm{f}}\boldsymbol{v}) + (\boldsymbol{p}_{2\mathrm{f}} - m_{2\mathrm{f}}\boldsymbol{v}),$$

thus,

$$(\boldsymbol{p}_{1\mathrm{i}} + \boldsymbol{p}_{2\mathrm{i}}) - \underbrace{(m_{1\mathrm{i}} + m_{2\mathrm{i}})\boldsymbol{v}}_{A_\mathrm{i}} = (\boldsymbol{p}_{1\mathrm{f}} + \boldsymbol{p}_{2\mathrm{f}}) - \underbrace{(m_{1\mathrm{f}} + m_{2\mathrm{f}})\boldsymbol{v}}_{A_\mathrm{f}} \,.$$

This is equivalent to (13.17) if and only if $A_\mathrm{i} = A_\mathrm{f}$ or

$$m_{1\mathrm{i}} + m_{2\mathrm{i}} = m_{1\mathrm{f}} + m_{2\mathrm{f}}.$$

This is the **conservation of the total mass**. We see that, for the conservation of total momentum, the masses of the particles in the collision are allowed to change, but the sum of the masses must stay the same. So, one could imagine that, during the collision, a part of particle 1 breaks off and sticks to particle 2. Mass and momentum would still be conserved.

Altogether, we have:

From the fact that, in classical mechanics,
- the momentum is given by $\boldsymbol{p} = m\boldsymbol{v}$,
- the Galilei transformation holds, and
- the total mass in a collision is conserved,

it follows that, if the **total momentum** in a collision is conserved in one inertial frame, it is also conserved in any other inertial frame.

Energy. For the energy, a similar statement holds. We restrict ourselves to *elastic* collisions in which the sum of the kinetic energies of the colliding particles (i. e., the total kinetic energy) before the collision is the same as after the collision. For a particle with mass m and velocity $\boldsymbol{v}$, in classical mechanics, the kinetic energy is given by $E_\mathrm{kin} = m\boldsymbol{v}^2/2$. During the collision proper, the kinetic energy changes, and may even vanish completely in one moment.

For Bob (in the primed reference frame), the conservation of the total kinetic energy reads as (with a notation analogous as that used with the momentum)

$$E'_\mathrm{kin,i} = E'_\mathrm{kin,f} \quad \text{or} \quad E'_\mathrm{kin,1i} + E'_\mathrm{kin,2i} = E'_\mathrm{kin,1f} + E'_\mathrm{kin,2f} \tag{13.21}$$

or

$$\frac{m_{1i}}{2}(\boldsymbol{u}'_{1i})^2 + \frac{m_{2i}}{2}(\boldsymbol{u}'_{2i})^2 = \frac{m_{1f}}{2}(\boldsymbol{u}'_{1f})^2 + \frac{m_{2f}}{2}(\boldsymbol{u}'_{2f})^2.$$

Using $\boldsymbol{u}' = \boldsymbol{u} - \boldsymbol{v}$ in $E'_{\text{kin}} = m\boldsymbol{u}'^2/2$, one gets the **transformation law for the kinetic energy**

$$E'_{\text{kin}} = \frac{m}{2}(\boldsymbol{u} - \boldsymbol{v})^2 = \frac{m}{2}\boldsymbol{u}^2 - m\boldsymbol{u}\boldsymbol{v} + \frac{m}{2}\boldsymbol{v}^2 = E_{\text{kin}} - \boldsymbol{p}\boldsymbol{v} + \frac{m}{2}\boldsymbol{v}^2.$$

Therefore, from (13.21), it follows that

$$\begin{aligned} &E_{\text{kin},1i} + E_{\text{kin},2i} - \underbrace{(\boldsymbol{p}_{1i} + \boldsymbol{p}_{2i})\,\boldsymbol{v}}_{A_i} + \underbrace{\frac{1}{2}\,(m_{1i} + m_{2i})\,\boldsymbol{v}^2}_{B_i} \\ &= E_{\text{kin},1f} + E_{\text{kin},2f} - \underbrace{(\boldsymbol{p}_{1f} + \boldsymbol{p}_{2f})\,\boldsymbol{v}}_{A_f} + \underbrace{\frac{1}{2}\,(m_{1f} + m_{2f})\,\boldsymbol{v}^2}_{B_f}. \end{aligned}$$

Because of momentum conservation, we have $A_i = A_f$, and because of the mass conservation, $B_i = B_f$. The corresponding expressions cancel each other out, and we are left with the conservation of the total kinetic energy for Alice (in the non-primed reference frame):

$$E_{\text{kin},1i} + E_{\text{kin},2i} = E_{\text{kin},1f} + E_{\text{kin},2f}.$$

Therefore, we can conclude:

From the fact that, for *elastic* collisions, in classical mechanics
- the kinetic energy is given by $E_{\text{kin}} = m\boldsymbol{v}^2/2$,
- the Galilei transformation holds,
- the total mass in a collision is conserved,
- and the total momentum in a collision is conserved,

it follows that, for *elastic* collisions, if the **total kinetic energy** is conserved in one inertial frame, it is also conserved in any other inertial frame.

As stated, the conservation of the total kinetic energy, however, only holds in the special case of an elastic collision. Another extreme case is the collision of two particles 1 and 2 with velocities $\boldsymbol{v}_1 = -\boldsymbol{v}_2$ that directly collide and stick together (as in Fig. 13.3). Here, the total kinetic energy obviously is not conserved; it is larger than zero before the collision and zero thereafter. The total mass and the total momentum, however, are conserved (the total momentum before and after the collision is zero).

We have shown the consistency of the conservation laws with the relativity principle only for the collision of two particles. The conservation of total mass and total momentum, however, holds for collisions of an arbitrary number of particles. The conservation of the total kinetic energy only holds for a particular class of collisions, namely, *elastic* collisions (here, also, for an arbitrary number of particles).

Bear in mind the delicate interplay of the conservation laws and the principle of relativity. A conservation law only makes sense if it holds in any inertial frame. But this is possible for the total momentum only if the total mass is conserved as well. For the conservation law of the total kinetic energy in the case of elastic collisions, this only holds if the total momentum and the total mass are conserved.

13.5.4 Theory of Relativity with Classical Momentum

We come back to special relativity, in which the Lorentz transformation plays the role of the Galilei transformation.

How does the classical momentum $p_{\mathrm{cl}} = mu$ transform in this case? The addition of velocities is now given by the Lorentzian addition of velocities (10.3). As in the preceding chapter, we suppose that the mass is invariant, i. e., the mass of a particle is the same in all inertial frames.

Let us restrict our discussion to one dimension. Thus, for the transformation law of the classical momentum, we get

$$p'_{\mathrm{cl}} = mu' = m\frac{u - v}{1 - uv/c^2} = \frac{p_{\mathrm{cl}} - mv}{1 - p_{\mathrm{cl}}v/(mc^2)},$$

which cannot be written nicely with the classical momentum p_{cl} in the unprimed reference frame.

Suppose again that momentum conservation (13.18) holds for Bob (in primed coordinates). With the transformation of the classical momentum, we get

$$m_{1\mathrm{i}}\frac{u_{1\mathrm{i}} - v}{1 - u_{1\mathrm{i}}v/c^2} + m_{2\mathrm{i}}\frac{u_{2\mathrm{i}} - v}{1 - u_{2\mathrm{i}}v/c^2} = m_{1\mathrm{f}}\frac{u_{1\mathrm{f}} - v}{1 - u_{1\mathrm{f}}v/c^2} + m_{2\mathrm{f}}\frac{u_{2\mathrm{f}} - v}{1 - u_{2\mathrm{f}}v/c^2}\,. \tag{13.22}$$

This formula cannot be cast into the form of a momentum conservation law for Alice. We have failed (for the time being). The momentum in special relativity cannot be given by $p_{\mathrm{cl}} = mu$. The reason for this, ultimately, is the relativistic addition of velocities, which causes the different denominators in the formula above.

13.5.5 Theory of Relativity

Now, we use the correct relativistic expressions for the energy E (see (13.8)), the momentum $\boldsymbol{p}$ (see (13.12)), and the respective Lorentz transformation (13.13).

Then, starting from the conservation of momentum for Bob (13.18), we get

$$\boldsymbol{p}_{1\mathrm{i}} + \boldsymbol{p}_{2\mathrm{i}} - \frac{v}{c^2}\left(E_{1\mathrm{i}} + E_{2\mathrm{i}}\right) = \boldsymbol{p}_{1\mathrm{f}} + \boldsymbol{p}_{2\mathrm{f}} - \frac{v}{c^2}\left(E_{1\mathrm{f}} + E_{2\mathrm{f}}\right).$$

In other words: if momentum and energy are conserved for Alice, then momentum is conserved for Bob.

Now, we start from the conservation of energy for Bob (13.19) and get

$$E_{1\mathrm{i}} + E_{2\mathrm{i}} - v\left(\boldsymbol{p}_{1\mathrm{i}} + \boldsymbol{p}_{2\mathrm{i}}\right) = E_{1\mathrm{f}} + E_{2\mathrm{f}} - v\left(\boldsymbol{p}_{1\mathrm{f}} + \boldsymbol{p}_{2\mathrm{f}}\right).$$

In other words: if momentum and energy are conserved for Alice, then energy is conserved for Bob.

If we combine these two findings and also take into account that, instead of having started with the conservation laws for Bob, we can also start with Alice, we get:

> If, in special relativity, for a collision of two point-like particles, energy and momentum are conserved for Alice, these quantities are also conserved for Bob.

We have made the calculations above for the special case of a collision between two particles. Indeed, the observation above is much more general. Note that, in special relativity, we do not have a separate conservation of momentum anymore. This is included in the conservation of energy and momentum

13.6 The Compton Effect

We finish this chapter with an interesting application of the relativistic energy and momentum, an experiment that was an important cornerstone in the development of special relativity, but also of quantum theory.

Introduction. In 1922, Arthur H. Compton investigated the scattering of monochromatic electromagnetic waves with a high energy (X-rays[17]) from free electrons at rest. Among other things, he observed that the frequency of the scattered wave was smaller that that of the incoming wave. The wave *experienced a change of frequency*, in which the amount of change depends on the direction of scattering.

This **Compton effect** is not comprehensible through the combination of electrodynamics and classical mechanics, because, according to these theories, the incoming wave would force the free electron to oscillate with the exact frequency of the incoming wave. The acceleration of the electron would then cause an emission of an

[17] X-rays, also known as Röntgen rays, after their discoverer, the first ever Nobel laureate, Wilhelm Conrad Röntgen.

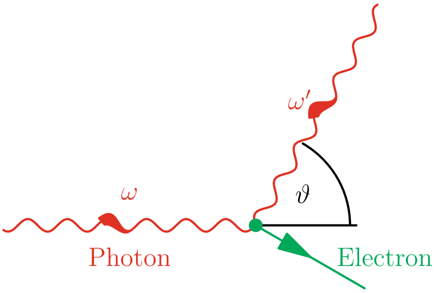


Fig. 13.9 Regarding the Compton effect

electromagnetic wave with exactly the same frequency as its oscillation. The scattered (or emitted) wave would therefore have exactly the same frequency as the incoming wave. In other words: The scattering would be elastic. This kind of scattering is called **Thomson scattering**.

To explain the Compton effect, we have to assume that light consists of a kind of particle, called a **photon**.[18] The Compton effect, in addition, is a relativistic effect. Using classical mechanics, one can derive a formula for the Compton effect. But this formula only works at relatively low photon frequencies. A formula that is in agreement with the experimental findings in the case of high frequencies can only be derived within the framework of the special theory of relativity. The Compton effect therefore breaks completely with classical physics: it needs quantum physics and special relativity for its explanation.

Setup and conservation laws. Figure 13.9 schematically shows the experiment. A photon from the beam of a monochromatic wave with frequency ω collides with an electron at rest. After the collision, the photon has the frequency ω' and, in comparison to earlier, its propagation direction is changed by an angle ϑ.

In the scattering process, the (relativistic) energy and the (relativistic) momentum must be conserved. Consider first the energy conservation and denote the energy of the photon before the scattering event with E_p and that of the electron with E_e. The related quantities after the scattering event are E'_p and E'_e, respectively. Energy conservation means that

$$E_\mathrm{p} + E_\mathrm{e} = E'_\mathrm{p} + E'_\mathrm{e}. \tag{13.23}$$

For the momentum, we denote the quantities in an analogous way. Momentum conservation means that

$$\boldsymbol{p}_\mathrm{p} + \boldsymbol{p}_\mathrm{e} = \boldsymbol{p}'_\mathrm{p} + \boldsymbol{p}'_\mathrm{e}. \tag{13.24}$$

The relation between the energies and momenta is given by the energy-momentum relation (13.14). In general, we have

[18] Photons ("light quanta") were introduced by Einstein, also in his *annus mirabilis* 1905, to explain the photoelectric effect, in which electrons are emitted by a solid when it is illuminated with light.

$$E^2 = \boldsymbol{p}^2c^2 + m^2c^4.$$

The mass of the photon vanishes, therefore, $E_\mathrm{p}^2 = \boldsymbol{p}_\mathrm{p}^2c^2$ or $E_\mathrm{p} = c\big|\boldsymbol{p}_\mathrm{p}\big|$. For the electron, we have $E_\mathrm{e}^2 = \boldsymbol{p}_\mathrm{e}^2c^2 + m_\mathrm{e}^2c^4$, where m_e is the mass of the electron.

We choose the inertial frame such that the electron is at rest before the scattering event, so that $\boldsymbol{p}_\mathrm{e} = 0$. All other momenta lie in a plane, which, by appropriate orientation of the coordinate system, is the xy-plane, and we can neglect the z-components of the momenta in what comes. We furthermore orient the x-axis such that the momentum of the incoming photon lies on the x-axis and both point in the same direction.

Energy and momentum of the photon. How do the energy and momentum of the photon after the scattering event depend on its frequency ω? For the energy, we already know the answer (see Sect. 13.1). Einstein's analysis of the photoelectric effect has shown that a photon that is part of a monochromatic light beam of frequency ω has an energy of $E = \hbar\omega$. The expression for the photon's momentum now follows from the energy-momentum relation $E = c|\boldsymbol{p}|$. We have $|\boldsymbol{p}| = E_\mathrm{p}/c = \hbar\omega/c$. Because of $\omega = c|\boldsymbol{k}|$, this can be written as $|\boldsymbol{p}| = \hbar|\boldsymbol{k}|$. The momentum of the photon points in the same direction as the wavevector, and therefore we have $\boldsymbol{p} = \hbar\boldsymbol{k}$. To summarize, we have the following situation:

"particle picture"	Relation	"Wave picture"
$E = c\lvert\boldsymbol{p}\rvert \left\{ \begin{matrix} \boldsymbol{p} \\ E \end{matrix} \right.$	$\boldsymbol{p} = \hbar\boldsymbol{k}$ $E = \hbar\omega$	$\left. \begin{matrix} \boldsymbol{k} \\ \omega \end{matrix} \right\} \omega = c\lvert\boldsymbol{k}\rvert$

The photon after the scattering event. It follows that

$$\begin{aligned} \boldsymbol{p}_\mathrm{p} &= \frac{\hbar\omega}{c}\begin{pmatrix}1\\0\end{pmatrix}, & \boldsymbol{p}_\mathrm{p}' &= \frac{\hbar\omega'}{c}\begin{pmatrix}\cos\vartheta\\ \sin\vartheta\end{pmatrix}, \\ \boldsymbol{p}_\mathrm{e} &= 0, & \boldsymbol{p}_\mathrm{e}' &= \text{unknown}. \end{aligned}$$

The momentum of the electron after the collision can be eliminated from the equations, as it is not of interest for us and is usually not accessible in the experiment. For the energies, by means of the energy-momentum relation, we have

$$\begin{aligned} E_\mathrm{p} &= c\big|\boldsymbol{p}_\mathrm{p}\big|, & E_\mathrm{p}' &= c\Big|\boldsymbol{p}_\mathrm{p}'\Big|, \\ E_\mathrm{e}^2 &= \boldsymbol{p}_\mathrm{e}^2c^2 + m_\mathrm{e}^2c^4 = m_\mathrm{e}^2c^4, & E_\mathrm{e}'^2 &= \boldsymbol{p}_\mathrm{e}'^2c^2 + m_\mathrm{e}^2c^4. \end{aligned} \tag{13.25}$$

The equation at the bottom right mutually relates the unknown energy and the unknown momentum of the electron after the collision and is the starting point for eliminating these quantities. We solve the conservation Eqs. (13.23) and (13.24) for E_e' and $\boldsymbol{p}_\mathrm{e}'$ (note that $\boldsymbol{p}_\mathrm{e} = 0$) and plug the result into the equation at the bottom left of (13.25). This results in

$$(E_\mathrm{p} + E_\mathrm{e} - E'_\mathrm{p})^2 - (\boldsymbol{p}_\mathrm{p} - \boldsymbol{p}'_\mathrm{p})^2 c^2 = m_\mathrm{e}^2 c^4.$$

Next, we calculate the squares of the expressions in parenthesis and get

$$(E_\mathrm{p}^2 + E_\mathrm{e}^2 + E_\mathrm{p}'^2 + 2E_\mathrm{p}E_\mathrm{e} - 2E_\mathrm{p}E'_\mathrm{p} - 2E_\mathrm{e}E'_\mathrm{p}) - (\boldsymbol{p}_\mathrm{p}^2 - 2\boldsymbol{p}_\mathrm{p}\boldsymbol{p}'_\mathrm{p} + \boldsymbol{p}_\mathrm{p}'^2)c^2 = m_\mathrm{e}^2 c^4.$$

With (13.25) and $E_\mathrm{e}^2 = m_\mathrm{e}^2 c^4$, this immediately becomes

$$E_\mathrm{p}E_\mathrm{e} - E_\mathrm{p}E'_\mathrm{p} - E_\mathrm{e}E'_\mathrm{p} + \boldsymbol{p}_\mathrm{p}\boldsymbol{p}'_\mathrm{p}c^2 = 0.$$

Now, we only have to plug in the expressions for the energies and the momenta. In particular, $\boldsymbol{p}_\mathrm{p}\boldsymbol{p}'_\mathrm{p} = (\hbar/c)^2\omega\omega' \cos\vartheta$ (see Fig. 13.9). Then,

$$m_\mathrm{e}c^2\hbar\omega - \hbar^2\omega\omega' - m_\mathrm{e}c^2\hbar\omega' + \hbar^2\omega\omega' \cos\vartheta = 0.$$

Rearranging and dividing by $m_\mathrm{e}c^2\hbar$ brings us to

$$\omega - \omega' = \frac{\hbar}{m_\mathrm{e}c^2}\omega\omega'(1 - \cos\vartheta) = \frac{\omega\omega'}{\omega_\mathrm{C}}(1 - \cos\vartheta), \tag{13.26}$$

where $\omega_\mathrm{C} = m_\mathrm{e}c^2/\hbar$ is the **Compton frequency**.[19] This is the result that we were looking for. To bring it into the usual form, we multiply it with $2\pi c/(\omega\omega')$ and note that, for the wavelength, $\lambda = 2\pi c/\omega$ holds. This finally brings us to the famous **Compton formula**

$$\lambda' - \lambda = \lambda_C \cdot (1 - \cos\vartheta), \tag{13.27}$$

which gives us the change of the photon's wavelength due to the scattering event.

Here, $\lambda_C = h/(m_\mathrm{e}c)$ is the **Compton wavelength** of the electron and the fundamental physical constant h is the Planck constant.

Energy/frequency change. We will discuss this result now, and the discussion will be based on frequencies, not wavelengths. Therefore, we have to rearrange Compton's formula again. We are interested in $E'_\mathrm{p}/E_\mathrm{p} = \omega'/\omega$, the energy/frequency of the photon after the scattering event in relation to its energy/frequency before the scattering event. Rearranging (13.26) gives us the *energy/frequency change*

$$\frac{E'_\mathrm{p}}{E_\mathrm{p}} = \frac{\omega'}{\omega} = \frac{1}{1 + \frac{\omega}{\omega_\mathrm{C}}(1 - \cos\vartheta)}\,. \tag{13.28}$$

[19] The energy of a photon with the Compton frequency ω_C is $\hbar\omega_\mathrm{C} = m_\mathrm{e}c^2$, i. e., equal to the rest energy of the electron. The Compton frequency is very large, about a million times larger than frequencies of electromagnetic waves in the visible region, and lies in the region between hard X-rays and γ-rays.

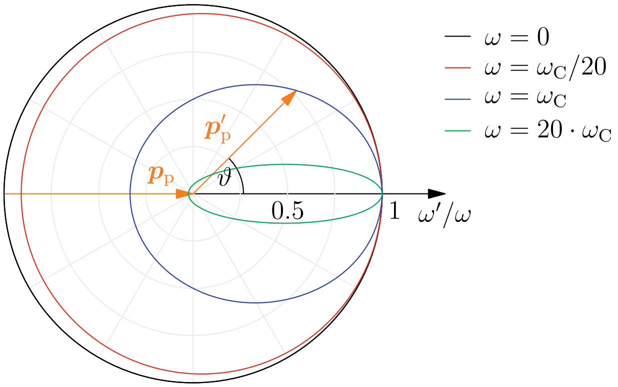


Fig. 13.10 Energy/frequency *change* ω'/ω of the photon for Compton scattering. The diagram is a polar diagram. The orange vectors, as an example, show an incoming photon with momentum $\boldsymbol{p}_p$ and energy ω, which is scattered into an angle ϑ and ends up with momentum $\boldsymbol{p}'_p$ and energy ω'

The value on the right side lies always in the interval (0, 1]. This is expected, because the electron gains energy from the collision. Therefore, the photon must lose energy, and this means that its frequency becomes smaller. For $\omega' = \omega$, we recover the classical limit, Thomson's result, when the energy of the photon does not change due to the scattering event.

Discussion. We discuss two special cases. For $\vartheta = 0$, the photon is not scattered at all. This is called *forward scattering*. Then, according to (13.28), $\omega' = \omega$, i. e., the photon's frequency does not change. For $\vartheta = \pi/2$, the photon is scattered back into the direction from which it came. This is called *backscattering*. Then, from (13.28), we have

$$\frac{\omega'}{\omega} = \frac{1}{1 + 2\frac{\omega}{\omega_C}}.$$

In this case, ω'/ω is always smaller than one (except in the classical limit $\omega \to 0$). The photon loses energy. The higher its frequency, the larger the amount of energy it loses. If the initial photon energy is equal to the electron's mass, we have $\omega = \omega_C$ and get $\omega'/\omega = 1/3$, so the photon loses 2/3 of its energy.

In the polar diagram in Fig. 13.10, the relative energy ω'/ω of the scattered photon as a function of the scattering angle ϑ is shown for three different photon energies and the classical limit ($\omega = 0$, black curve). The red curve is for $\omega = \omega_C/20$, the green curve for $\omega = \omega_C$ and the blue curve for $\omega = 20\,\omega_C$. Independent of its energy, the photon does not lose energy in forward scattering. Thus, the larger the scattering angle ϑ, the smaller the energy after the scattering (or the larger the energy loss). Due to the fact that $\omega = c\left|\boldsymbol{p}_p\right|$, the magnitude of the momentum is proportional to the frequency. Therefore, we can directly plot the momentum of the photon before and after the scattering into the polar diagram.

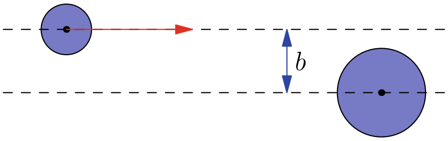


Fig. 13.11 Definition of the impact parameter b. The "target" is at rest before the collision

For $\omega = \omega_C/20$, the photon loses very little energy— independent of the scattering direction. The case $\omega \ll \omega_C$ is the limiting case in which Thomson's result (black curve) is valid. For $\omega = \omega_C$, one sees the energy loss of $2/3$ in backward scattering. And for $\omega = 20\,\omega_C$, the photon for scattering angles of $\vartheta > 20°$ loses almost all of its energy.

Differential scattering cross section. In addition to the energy/frequency change as a function of the scattering direction in (13.28), the *differential scattering cross section* is an important quantity.

Consider two billiard balls: one is at rest and the other moves uniformly toward the first one, whereupon they collide (see Fig. 13.11).[20] Then, the directions in which both billiard balls move after the collision are uniquely given by the *impact parameter b* before the collision. This is the nearest distance of the line along which the center of mass of the moving ball moves, from the center of mass of the billiard ball at rest. In the case at hand, the scattering angle of the photon would be a function of the impact parameter: $\vartheta = \vartheta(b)$, and we have $\vartheta(0) = \pi$, because, if the moving billiard ball moves exactly toward the resting billiard ball, the former will be backscattered.

In quantum theory, the situation is different. If the billiard balls were quantum objects,[21] then, even if the moving billiard ball were to move exactly toward the center of the resting billiard ball, all scattering angles ϑ would be possible. Indeed, in an exact repetition of the scattering experiment, we would find all different scattering angles. The probability distribution of the scattering angle (i. e., how often the different angles would be found in the repeated experiment), however, would be fixed. And this probability distribution is called the **differential scattering cross section**. The differential scattering cross section is a function of the scattering angle ϑ and gives us the probability that the photon is scattered in the direction ϑ.

In 1904, Joseph J. Thomson, the discoverer of the electron, calculated this scattering cross section for the scattering of a monochromatic wave at an electron at rest—exactly the case that we consider here. At that time, there was neither the special theory of relativity nor quantum theory, and his calculation was based on classical physics. For the scattering cross section, he got a direction dependence of $1 + \cos^2\vartheta$. Forward scattering ($\vartheta = 0$) and backscattering ($\vartheta = \pi$) are equally probable in classical physics and more probable than the scattering in any other direction. The function is shown for arbitrary ϑ by the peanut-shaped black curve of Fig. 13.12. For photon frequencies within the range of the Compton frequency and above, this scattering cross section, however, is no longer correct.

[20] The balls do not rotate before the scattering event.

[21] They would have to be much much lighter to require quantum theory for a description.

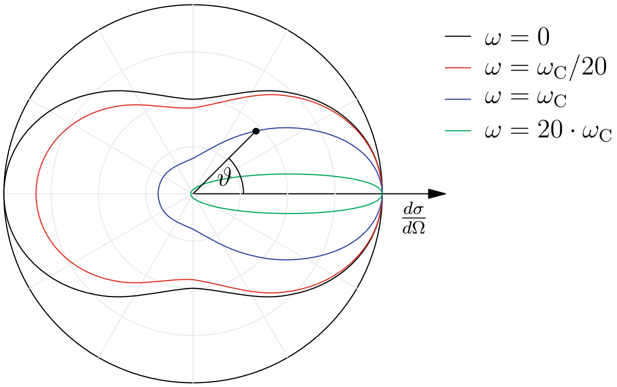


Fig. 13.12 Klein-Nishina differential scattering cross section for different photon energies $\hbar\omega$. The photon comes from the left. The diagram is a polar diagram and the angle ϑ the scattering angle. For instance, the probability that a photon with frequency $\omega = \omega_C$ becomes scattered in the direction ϑ is given by the black dot

Oskar Klein and Yoshio Nishina, in 1929, carried out the analogous (but much more difficult) calculation on the basis of quantum electrodynamics.[22] The result holds for all photon frequencies. In Fig. 13.12, the direction dependence of the differential scattering cross section for different frequencies ω is shown. The black curve ($\omega = 0$) is Thomson's result, the classical limit.

For small energies (red curve), the Klein-Nishina result is still very similar to Thomson's result and forward scattering and backscattering are of almost equal size. For larger energies (blue curve, $\omega = \omega_C$), forward scattering becomes more important. For very large energies (green curve), forward scattering predominates and the scattering almost exclusively goes into a small cone pointing in the forward direction.[23]

[22] Quantum electrodynamics is the "quantum version" of the special theory of relativity and electrodynamics.

[23] You may counter that the area inside of the objects that are created by rotating the curves in Fig. 13.12 around the axis should be one, otherwise, it cannot be interpreted as a probability density. The point here is that the missing probability corresponds to the cases in which no scattering at all happens.

Chapter 14
Electrodynamics

14.1 Transformation of Charges and Fields

At the end of this book about special relativity, we will take a short excursion into electrodynamics. You already know that electrodynamics "per construction" keeps its shape (is form-invariant) under a Lorentz transformation, so it is also correct at large velocities. Due to the fact that, in mechanics, upon a change from one inertial frame to another, quantities like the energy and the momentum transform, we also have to reckon with the central quantities of electrodynamics: the charge, the current and the electric and magnetic fields. But the transformation laws for these quantities can be found without much thinking. One must apply the Lorentz transformation to the laws of electrodynamics (the so-called Maxwell equations) and then bring them back to the standard form. The transformation equations for the quantities of electrodynamics can then be read off. This procedure is conceptually simple, but the calculations are tedious. Moreover, one does not learn much about physics. Therefore, we go a different way here.

Electric charge. We start with the **charge**. *The charge remains invariant upon a Lorentz transformation, which is* $q' = q$. This can be seen by the fact that atoms and molecules are neutral, even though the charged carriers in these systems have very different velocities, which also are comparable with the speed of light. If one excites, e. g., an atom, then the electron gets a higher velocity. If this were to change its charge, the electron's charge could no longer exactly compensate the charge of the atomic nucleus and the atom would suddenly no longer be neutral.

Another example is a solid metallic body in which ions oscillate around their equilibrium positions while the "conduction electrons" move freely. From thermodynamics, you know that the average kinetic energy $\frac{m}{2}\left\langle v^2\right\rangle$ of a particle equals $\frac{3}{2}k_B T$ (k_B is Boltzmann's constant and T the absolute temperature). If you increase the temperature, the average velocity of the electrons will grow much faster than that of the ions, because the latter have a much larger mass. Therefore, a temperature change also would cause the charge neutrality to disappear. None of this is observed, so we conclude that the electric charge is invariant.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_14

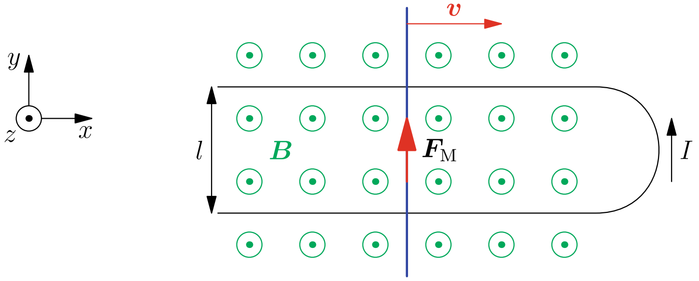


Fig. 14.1 An induction experiment

Electric and magnetic field. Next, we discuss the **fields**. Einstein starts his paper [Einstein05a], which introduces the special theory of relativity, with the words[1]:

> It is known that Maxwell's electrodynamics—as usually understood at the present time—when applied to moving bodies, leads to asymmetries that do not appear to be inherent in the phenomena. Take, for example, the reciprocal electromagnetic action of a magnet and a conductor. The observable phenomenon here depends only on the relative motion of the conductor and the magnet, whereas the customary view draws a sharp distinction between the two cases in which either that or the other of these bodies is in motion. For if the magnet is in motion and the conductor at rest, there arises in the neighborhood of the magnet an electric field with a certain definite energy, producing a current at the places where parts of the conductor are situated. But if the magnet is stationary and the conductor in motion, no electric field arises in the neighborhood of the magnet. In the conductor, however, we find an electromotive force, to which in itself there is no corresponding energy, but which gives rise – assuming equality of relative motion in the two cases discussed – to electric currents of the same path and intensity as those produced by the electric forces in the former case.[2]

We see that Einstein, in particular, by analyzing the following **induction phenomenon**, developed the central idea behind the special theory of relativity (see Fig. 14.1). Two fixed conducting metal rods are arranged parallel to each other, along the x-direction, with a mutual distance l. The metal rods are connected. A further metal rod lies on the other two, along the y-direction, and can be moved freely in the x-direction. Between the fixed metal rods, there is a homogeneous magnetic field $\boldsymbol{B} = (0, 0, B_0)$ with $B_0 > 0$. The movable metal rod is moved with constant velocity $\boldsymbol{v} = (v, 0, 0)$, $v > 0$, over the fixed metal rods. Then, there is a (magnetic) Lorentz force[3] $\boldsymbol{F}_{\text{mag}} = q\boldsymbol{v} \times \boldsymbol{B} = -e_0\boldsymbol{v} \times \boldsymbol{B}$, or $\boldsymbol{F}_{\text{mag}} = (0, F_0, 0)$ with $F_0 = e_0 v B_0$ ($e_0 > 0$ is the elementary charge), acting on the free electrons in the movable metal rod. The magnetic force causes an electric current to flow through the circuit formed

[1] The text is the author's translation of the original German text.

[2] By "electromotive force", Einstein refers to the Lorentz force.

[3] The term *Lorentz force* in literature is not used in a unique way. The Lorentz force is a force that acts on a moving charge. Some authors use the term for the force caused by the magnetic field, some others for the force caused by the magnetic and the electric field. We write "(magnetic) Lorentz force" and restrict it to the influence of the magnetic field. Note that neither is the space component of a four-vector.

by the metal rods. The metal rods stay electrically neutral, so there is no electric field in the experiment (in the stationary case).

Suppose that we described the experiment in Alice's inertial frame, where the movable metal rod moves with the velocity v (we assume that the standard configuration prevails). Now, we change to Bob's inertial frame, where the same metal rod is at rest. How does Bob explain the experiment? Also for him, there is an electric current flowing through the metal rod (the metal becomes warm). But what is the reason for this current? The electrons in the metal rod do not move in the x'-direction, therefore, there is no (magnetic) Lorentz force. The only possible cause of the current would be an electric field. For Bob, *there must be an electric field* that acts on the electrons in the metal rod.

Lorentz transformation of the fields. If Einstein's principle of relativity holds and one requires that the Maxwell equations be form-invariant upon Lorentz transformations, the electric and magnetic fields transform when we go from Alice's to Bob's inertial frame. For the standard configuration (Bob's coordinates for Alice are: $x = vt$, $y = z = 0$), the transformation law is

$$\begin{aligned} E'_{x'} &= E_x \;, & B'_x &= B_x \;, \\ E'_{y'} &= \gamma_v(E_y - vB_z) \;, & B'_{y'} &= \gamma_v(B_y + \frac{v}{c^2}E_z) \;, \\ E'_{z'} &= \gamma_v(E_z + vB_y) \;, & B'_{z'} &= \gamma_v(B_z - \frac{v}{c^2}E_y) \;. \end{aligned} \tag{14.1}$$

Here, $\boldsymbol{E} = (E_x, E_y, E_z)$ and $\boldsymbol{B} = (B_x, B_y, B_z)$ are the fields for Alice and $\boldsymbol{E}' = (E'_{x'}, E'_{y'}, E'_{z'})$ and $\boldsymbol{B}' = (B'_{x'}, B'_{y'}, B'_{z'})$ those for Bob. The field components parallel to the relative velocities of Bob and Alice are the same for both, but the components perpendicular to it are different.[4] We won't derive these formulas here (above, we mentioned how the derivation is performed), but we will discuss special cases in Sect. 14.2.

Remember the invariants of spacetime (9.2), the four-wavevector (12.16) and energy and momentum (121)? These are due to the Lorentz transformation and, therefore, we have an invariant here as well. Actually, we even have two invariants,[5] which are

$$\boldsymbol{E}^2 - c^2\boldsymbol{B}^2 \quad \text{and} \quad \boldsymbol{E} \cdot \boldsymbol{B} \;.$$

The first invariant tells us that, if, in one inertial frame, $|\boldsymbol{E}| > c|\boldsymbol{B}|$, this is also the case in all other inertial frames. And the second tells us that the angle between $\boldsymbol{E}$ and $\boldsymbol{B}$ is the same in all inertial frames. If $\boldsymbol{E} \perp \boldsymbol{B}$ in one inertial frame, this is also the case in all other inertial frames.

[4] Note that we have exactly the reverse situation with length contraction, which is only present in the longitudinal but not in the transversal direction.

[5] To satisfy your curiosity: this is due to the fact that the electric and the magnetic field vectors together form a kind of two-dimensional four-vector, a so-called second rank four-tensor.

But let us come back to the Lorentz transformation. In the situation at hand (Fig. 14.1), the fields for Alice are

$$\boldsymbol{E} = 0\,, \quad \boldsymbol{B} = (0, 0, B_0) \text{ with } B_0 > 0\,.$$

Those for Bob, according to (14.1), become

$$\boldsymbol{E}' = (0, -\gamma_v v B_0, 0)\,, \quad \boldsymbol{B}' = (0, 0, \gamma_v B_0)\,.$$

From Bob's perspective, there is indeed an electric field. For small velocities v with $\gamma_v \approx 1$, it causes the electric force $F_{\text{el}} = qE = -e_0 E \approx e_0 v B_0$ on the electrons in the metal rod. For Bob, this electric force F_{el} for small velocities has the same magnitude and the same cause as the (magnetic) Lorentz force F_{m} for Alice.

Electromagnetic field. Equation (14.1) shows that, upon change of the inertial frame, one can transform electric fields into magnetic fields, and vice versa (at least partly). For this reason, it makes sense to talk about the **electromagnetic field**, and not separately about the *electric* and the *magnetic fields*. There is only one field, the electromagnetic field, which presents itself in a different way in different inertial frames. Note that, with the same argument, one also should include the action of the magnetic and the electric fields in the Lorentz force, which becomes $\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$.

We again can adapt Minkowski's words from the end of Sect. 11.3 and say:

> The views of the *electric* and the *magnetic fields* that I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth, the *electric field* by itself, and the *magnetic field* by itself, are doomed to fade away into mere shadows, and only a kind of union of the two, the *electromagnetic field* will preserve an independent reality.

14.2 Electrodynamics and Spacetime Effects

In this section, we discuss two Gedanken experiments that show that the electromagnetic field must transform under a Lorentz transformation. Otherwise, it would not be consistent with the kinematic effects of the special theory of relativity (relativity of simultaneity, length contraction, time dilation).

14.2.1 The Charged Capacitor

The first Gedanken experiment deals with a vary large **charged capacitor** (a parallel-plate capacitor; see Fig. 14.2). In Alice's inertial frame, the capacitor is at rest. It is charged with the surface charge density σ (in Coulombs per square meter), and, therefore, in its interior, a homogeneous electric field

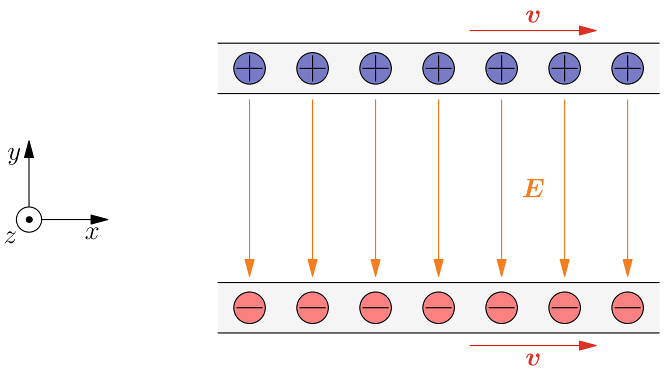


Fig. 14.2 A charged capacitor, moving with velocity v

$$E_0 = \frac{1}{\epsilon_0}|\sigma| \tag{14.2}$$

perpendicular to the plates. Here, ϵ_0 is the *vacuum permittivity*, a fundamental physical constant.[6] Bob moves relative to Alice and in parallel to the capacitor with velocity v. For Bob, the capacitor plates move, and therefore they are shorter by a factor of γ_v^{-1} than they are for Alice. But the charge is an invariant, therefore, the surface charge density σ' must be larger than for Alice: it is $\sigma' = \gamma_v \sigma$. For that reason, the electric field must also be larger for Bob than for Alice: $E' = \gamma_v E_0$. The exact same result that we got from (14.1).

Exercise 60: For the situation in Fig. 14.2, and using (14.1), from the electric field $\boldsymbol{E}$ in Alice's inertial frame, derive the magnetic field $\boldsymbol{B}$ that exists in Bob's inertial frame.

14.2.2 The Current-Carrying Wire

Consider a metal wire with free electrons (that move freely in the wire) and fixed positive charges (which are at rest relative to the wire). To make the following arguments as simple as possible, we will make the assumption that all electrons are equidistant and the same holds for the positive charges. This implies that all electrons move with the same velocity, as the positive charges do. Obviously this assumption is unrealistic. Still, it is not against the laws of physics, and therefore it can be used as a Gedanken experiment to draw valid conclusions.

Alice's point of view. First, we describe the situation in Alice's inertial frame, where *the wire and the positive charges are at rest* (see Fig. 14.3 on the left side).

[6] By the way: there are different unit systems in electrodynamics. Depending on the unit system, the formulas have different prefactors. We use the SI system.

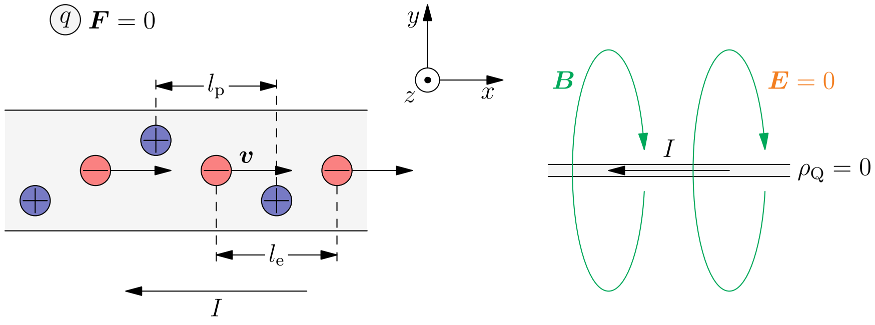


Fig. 14.3 Alice's perspective: current-carrying wire at rest, accompanied by resting positive charges and moving electrons. Furthermore, a resting test charge

Let l_e be the average distance between the electrons and l_p that between the positive charges. With $e_0 > 0$ being the elementary charge, we then have the (one-dimensional) charge densities $\rho_{\mathrm{Q,e}} = -e_0/l_\mathrm{e}$ and $\rho_{\mathrm{Q,p}} = e_0/l_\mathrm{p}$.

The wire, for Alice, has to be neutral, therefore, $\rho_{\mathrm{Q,p}} + \rho_{\mathrm{Q,e}} = 0$; and from this, it follows that $l_\mathrm{e} = l_\mathrm{p}$. There is **no electric field** outside the wire. We will use the density $\rho_{\mathrm{Q},0} := \rho_{\mathrm{Q,p}} = -\rho_{\mathrm{Q,e}} > 0$ of the positive charges in the resting wire as a reference to compare charge densities.

Now, suppose that the electrons move to the right, with an average velocity $v > 0$ (see Fig. 14.3 on the right side). Then, there is an **electric current**. An electric current is always given by the charge density times the velocity of the charged particles, $I = \rho_\mathrm{Q} v$. Here, the electric current is caused by the moving electrons, pointing in the negative x-direction, and is given by

$$I_0 = \rho_{\mathrm{Q,e}} v = -\rho_{\mathrm{Q},0} v \ .$$

This electric current causes a **magnetic field** outside the wire. The form of this field is calculated in high school lessons in electricity. The field lines are concentric circles around the wire. The x-component and the radial component of the magnetic field vanish completely. Only the tangential component does not vanish, and its magnitude depends on the distance r from the wire: the farther away from the wire, the smaller it is. The tangential component is given by

$$B = \frac{\mu_0}{2\pi} \frac{I}{r} \ . \tag{14.3}$$

Here, μ_0 is the *vacuum permeability*, another fundamental physical constant. It is related to the *vacuum permittivity* ϵ_0 via $\epsilon_0 \mu_0 = 1/c^2$, and therefore only two of the three fundamental physical constants c, ϵ_0 and μ_0 are independent.

In the case at hand, we have

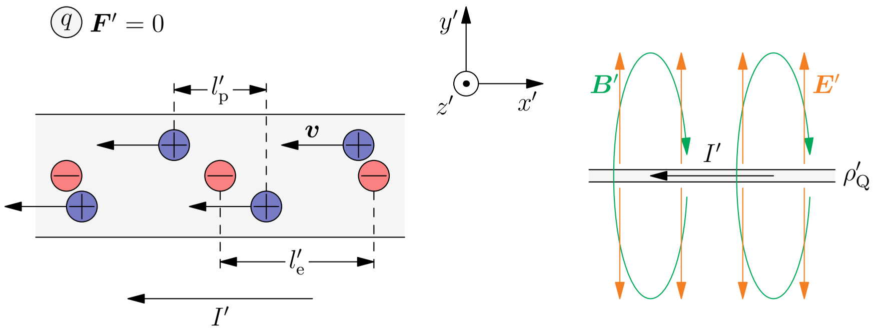


Fig. 14.4 Bob's perspective: moving current-carrying wire with comoving positive charges and electrons at rest. Furthermore, a test charge comoving with the wire

$$B = \frac{\mu_0}{2\pi}\frac{I_0}{r} = \frac{1}{2\pi\epsilon_0}\frac{1}{c^2}\frac{I_0}{r} = \frac{1}{2\pi\epsilon_0}\frac{\rho_{Q,0}}{r}\frac{v}{c^2}\,. \tag{14.4}$$

Bob's point of view. Now, we go to Bob's inertial frame, where *the electrons are at rest* and the positive charges move with velocity $-v < 0$ in the negative x'-direction (see Fig. 14.4 on the left side).

Due to the fact that the positive charges move, the distance between them is length-contracted, and Bob measures a charge density of $\rho'_\mathrm{Q,p} = e_0/l'_\mathrm{p}$, where $l'_\mathrm{p} = l_\mathrm{p}/\gamma_v$. So,

$$\rho'_\mathrm{Q,p} = \gamma_v \rho_\mathrm{Q,p} = \gamma_v \rho_\mathrm{Q,0}\,. \tag{14.5}$$

The charge density of the positive charges is larger for Bob than for Alice. The contrary holds for the charge density of the electrons, because the distance between these is length-contracted for Alice. Therefore, $l'_\mathrm{e} = \gamma_v l_\mathrm{e}$ and

$$\rho'_\mathrm{Q,e} = \gamma_v^{-1}\rho_\mathrm{Q,e} = -\gamma_v^{-1}\rho_\mathrm{Q,0}\,.$$

The charge density of the negative charges is smaller in magnitude for Bob than for Alice.

As a consequence, for Bob, the wire is *not electrically neutral* anymore: it has a positive charge density of

$$\rho'_\mathrm{Q} = \rho'_\mathrm{Q,p} + \rho'_\mathrm{Q,e} = (\gamma_v - \gamma_v^{-1})\rho_\mathrm{Q,0} = \gamma_v\frac{v^2}{c^2}\cdot\rho_\mathrm{Q,0} > 0\,.$$

Again, usually already in high school, we learn that a straight charged wire with a (linear) charge density of ρ_Q (in Coulombs per meter) creates an **electric field** in the radial direction which points away or toward the wire if the charge density is positive or negative, respectively, and which has the magnitude

$$E = \frac{1}{2\pi\epsilon_0}\frac{\rho_\mathrm{Q}}{r}\,. \tag{14.6}$$

Hence, due to the non-vanishing (net) charge $\rho'_{\rm Q}$, for Bob, there is an **electric field** outside the wire, with a radial component given by

$$E' = \frac{1}{2\pi\epsilon_0}\frac{\rho'_{\rm Q}}{r'} = \gamma_v \frac{v^2}{c^2}\cdot\frac{1}{2\pi\epsilon_0}\frac{\rho_{Q,0}}{r'} \ . \tag{14.7}$$

Here, we used $r' = r$, because the radial direction is perpendicular to the relative motion of Alice and Bob, and therefore there is no length contraction for the radial direction. The charge density for Bob is positive, therefore, the electric field points away from the wire.

In Bob's inertial frame, the electrons do not move, but the positive charges do. These cause an electric current

$$I' = \rho'_{\rm Q,p} v = \gamma_v \rho_{\rm Q,0} v = \gamma_v I_0$$

in the negative x'-direction. Therefore, the electric current is larger for Bob than for Alice. According to (14.3), this current causes a **magnetic field** of strength

$$B' = \frac{1}{2\pi\epsilon_0}\frac{1}{c^2}\frac{I'}{r'} = \gamma_v\frac{1}{2\pi\epsilon_0}\frac{1}{c^2}\frac{I}{r} = \gamma_v B \tag{14.8}$$

outside the wire.

Lorentz transformation. If we compare (14.4) to (14.7), we see that

$$E' = \gamma_v v B \quad \text{and} \quad B' = \gamma_v B \ .$$

This confirms (14.1) for the special case at hand. For instance, for Alice, on the x-y-plane for $y < 0$, the fields are

$$\begin{aligned} E_x &= 0 \ , & B_x &= 0 \ , \\ E_y &= 0 \ , & B_y &= 0 \ , \\ E_z &= 0 \ , & B_z &= B \ . \end{aligned}$$

where $B > 0$. According to (14.1), Bob then has

$$\begin{aligned} E'_{x'} &= 0 \ , & B'_{x'} &= 0 \ , \\ E'_{y'} &= -\gamma_v v B_z = -\gamma_v v B \ , & B'_{y'} &= 0 \ , \\ E'_{z'} &= 0 \ , & B'_{z'} &= \gamma_v B \ , \end{aligned}$$

exactly as expected from our example with the current-carrying wire.

Lorentz force. Consider again Alice's inertial frame (see Fig. 14.3). Suppose that, outside of the wire, there is a test charge $q > 0$ at rest. Because the test charge is at rest, there can be no (magnetic) Lorentz force acting on it, and because of the neutrality of the wire, there is also no electric field. Therefore, the electromagnetic field exerts no force on the test charge.

For Bob, in his inertial frame, the test charge moves with velocity v in the negative x'-direction (exactly as the electrons in the wire do). The electric field E' in Bob's inertial frame causes an electric force $F'_{\text{el}} = qE'$ on the test charge in the radial direction, away from the wire. There is also a (magnetic) Lorentz force because of the non-vanishing magnetic field and the fact that the test charge moves. Hence, we have the force $F'_{\text{mag}} = qvB'$, which also points in the radial direction, towards the wire. Due to the fact that, for Alice, the test charge does not move, it must also not move in the radial direction for Bob. Therefore, for Bob, the electric and the (magnetic) Lorentz forces on the test charge must compensate exactly. For this reason, **in the expression for the force, only the combination** $E' + vB'$ **must appear** (or, with vectors, $\boldsymbol{E}' + \boldsymbol{v} \times \boldsymbol{B}'$).[7]

Conclusion. Let us go back one step. What did we do in this section? First, (for a special case), we determined the transformation of charge and current between two inertial frames. We needed the invariance of the charge and the length contraction for that. Then, we were able to calculate the electric and magnetic fields in both inertial frames. Only the formulas $B = (\mu_0/2\pi)(I/r)$ and $E = (1/2\pi\epsilon_0)(\rho_Q/r)$ were needed. These follow directly from Maxwell's equations and are valid in all inertial frames. Comparison of the fields then gave us a confirmation of the transformation law (14.1). Therefore, this transformation law and the Lorentz transformation for space and time are consistent (at least, for the considered example).

14.2.3 The Four-Vector of the Current Density

In Sect. 14.2.2, using the simple example of equally spaced electric charges, we have seen that a charge density ρ that is at rest for Alice, according to (14.5), becomes the charge density $\rho' = \gamma_v \rho$ for Bob. Here, Bob moves with velocity v relative to Alice and so do the charges relative to Bob (but in the opposite direction). The charge density due to the moving charges is higher than for the same charges at rest. The simple reason for this is length contraction.

The **current density** j associated with a charge density ρ moving with velocity v is defined, in general, by $j = v\rho$. Therefore, Bob, as a consequence of the moving charges, sees an electric current with the current density $j' = v\rho'$.

Suppose now that the charge density ρ' and the current density $j' = \rho'$ for Bob are given and we want to calculate the respective values for Alice. The charges are at rest for Alice, and therefore, there is no current for her, hence, $j = 0$. We can get this by combining ρ' and $j' = v\rho'$ in the form $j = \alpha(j' - v\rho')$, where α is to be determined. But this looks like a Lorentz transformation, with j' playing the role of x' and ρ' that of t' (the other way around is not possible, because, in three dimensions,

[7] Only the combination $\gamma_v q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$ transforms like the space component of a four-vector, so, in special relativity, it is actually this combination that we should call the Lorentz force.

the current density is a vector while the charge density is not). If this were true, we would have $\alpha = \gamma_v$ and

$$\begin{aligned} \rho &= \gamma_v \cdot (\rho' - (v/c^2) j') , \\ j &= \gamma_v \cdot (j' - v\rho') . \end{aligned} \tag{14.9}$$

Indeed, from $\rho' = \gamma_v \rho$ and $j' = v\rho'$, we get ρ and $j = 0$. We conclude that

$$(c\rho, \boldsymbol{j})$$

is the **four-vector of the current density**.

It is a pleasure for us one again to give the stage to Herrmann Minkowski (see end of Sect. 11.3) and ask him for a statement. He says:

> The views of *electric current* and *electric density* that I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth, electric current by itself, and electric density by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.

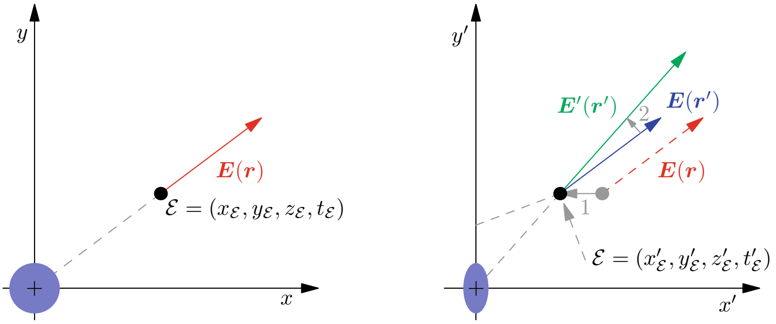
14.3 Electromagnetic Field of a Moving Point Charge

Transformation of fields. In an electromagnetic field, the quantities $\boldsymbol{E}$ and $\boldsymbol{B}$ depend, in general, on location and time. Suppose that, from Alice's point of view, we have the fields $\boldsymbol{E}(x, y, z, t)$ and $\boldsymbol{B}(x, y, z, t)$. What do these fields look like for Bob? To answer this question, it is not sufficient to just plug these quantities into (14.1), because this only transforms the vectors $\boldsymbol{E}$ and $\boldsymbol{B}$, but not the position and time of these vectors (i. e., the coordinates of the events where these vectors are pinned in spacetime). Equation (14.1) transforms the function $\boldsymbol{E}(x, y, z, t)$ into $\boldsymbol{E}'(x, y, z, t)$, but Bob needs $\boldsymbol{E}'(x', y', z', t')$.

The Lorentz transformation (14.1) holds for electromagnetic field vectors *at the same event* in spacetime and this event has different spacetime coordinates for Alice and Bob.

Let $\mathcal{E}$ be a certain event[8] in spacetime. Suppose that $\boldsymbol{E}(\mathcal{E})$ and $\boldsymbol{B}(\mathcal{E})$ are electric and magnetic field, respectively, at $\mathcal{E}$ for Alice. Further assume that the standard configuration prevails, which means that the Lorentz transformation of the coordinates is given by (11.9). Then, the electric $\boldsymbol{E}'(\mathcal{E})$ and the magnetic field $\boldsymbol{B}'(\mathcal{E})$ at $\mathcal{E}$, for Bob, are given by (14.1):

[8] We use the calligraphic $\mathcal{E}$ in order to be able to distinguish the event $\mathcal{E}$ from the magnitude E of the electric field.



y
$\boldsymbol{E}(\boldsymbol{r})$
$\mathcal{E} = (x_\mathcal{E}, y_\mathcal{E}, z_\mathcal{E}, t_\mathcal{E})$
x
y'
$\boldsymbol{E}'(\boldsymbol{r}')$
2
$\boldsymbol{E}(\boldsymbol{r}')$
$\boldsymbol{E}(\boldsymbol{r})$
1
$\mathcal{E} = (x'_\mathcal{E}, y'_\mathcal{E}, z'_\mathcal{E}, t'_\mathcal{E})$
x'

Fig. 14.5 Transformation of an electric field of a point charge at rest to a moving inertial frame. Left: The red vector is the electric field for Alice. Right: Step 1 in the transformation transforms the coordinates of the event $\mathcal{E}$ to Bob's coordinates. This moves the electric field vector to these new coordinates, but leaves the field vector untouched, resulting in the blue vector. Step 2, the Lorentz transformation of the fields, eventually leads to the green vector, the electric field for Bob. Note that the figure shows projections of the event $\mathcal{E}$ into the shown plane

$$\begin{aligned}
E'_x(\mathcal{E}) &= E_x(\mathcal{E}) \,, & B'_x(\mathcal{E}) &= B_x(\mathcal{E}) \,, \\
E'_y(\mathcal{E}) &= \gamma_v \left[E_y(\mathcal{E}) - vB_z(\mathcal{E})\right] \,, & B'_y(\mathcal{E}) &= \gamma_v \left[B_y(\mathcal{E}) + \frac{v}{c^2}E_z(\mathcal{E})\right] \,, \\
E'_z(\mathcal{E}) &= \gamma_v \left[E_z(\mathcal{E}) + vB_y(\mathcal{E})\right] \,, & B'_z(\mathcal{E}) &= \gamma_v \left[B_z(\mathcal{E}) - \frac{v}{c^2}E_y(\mathcal{E})\right] \,.
\end{aligned}$$

If we have an expression $\boldsymbol{E}(x, y, z, t)$ for a field for Alice, the coordinates (x, y, z, t) represent a particular event $\mathcal{E}$. The coordinates (x', y', z', t') of this event $\mathcal{E}$, for Bob, are given by the Lorentz transformation (11.9). This results in the following general **recipe** for the transformation of the fields (see Fig. 14.5 for the special case of the field of a point charge).

Suppose the electric and the magnetic field

$$\boldsymbol{E}(\boldsymbol{r}, t) = \begin{pmatrix} E_x(x, y, z, t) \\ E_y(x, y, z, t) \\ E_z(x, y, z, t) \end{pmatrix} \,, \quad \boldsymbol{B}(\boldsymbol{r}, t) = \begin{pmatrix} B_x(x, y, z, t) \\ B_y(x, y, z, t) \\ B_z(x, y, z, t) \end{pmatrix}$$

are given for Alice. To get the respective fields for Bob, we have to carry out *two steps*:

- **Step 1**: Apply the Lorentz transformation to the coordinates, i.e., perform the replacement $x \to x = \gamma_v \cdot (x' + vt')$, etc., in the expressions for $\boldsymbol{E}(\boldsymbol{r}, t)$ and $\boldsymbol{B}(\boldsymbol{r}, t)$. This gives the functions $\boldsymbol{E}(x', y', z', t')$ and $\boldsymbol{B}(x', y', z', t')$.
- **Step 2**: Then, plug these functions into the Lorentz transformation for the electromagnetic field (14.1). This yields the final result

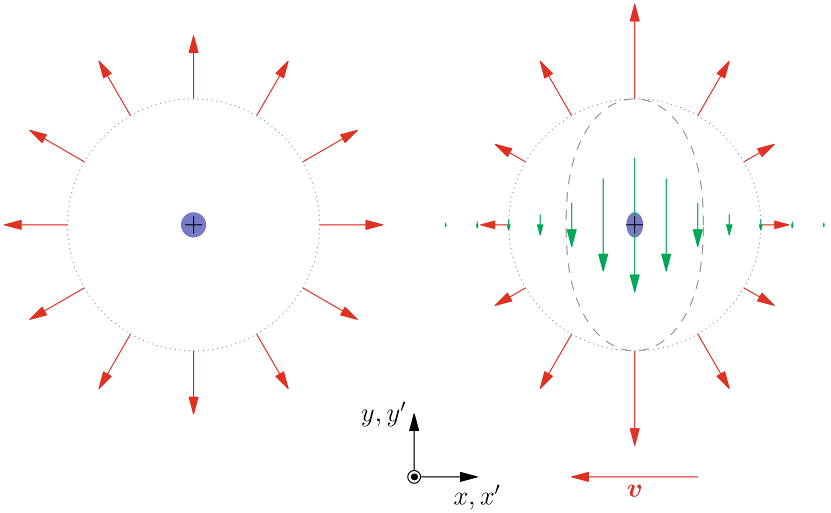


Fig. 14.6 Inertial point charge and the electromagnetic field caused by it. Left: For Alice, in the rest frame of the test charge. The red arrows are the electric field vectors on the dotted circle. Right: For Bob. Again, the red arrows are the electric field vectors on the dotted circle. On the dashed curve, the electric field has constant magnitude. The green arrows are the magnetic field on the line $(x, 1, 1)$

$$\boldsymbol{E}'(\boldsymbol{r}', t') = \begin{pmatrix} E'_{x'}(x', y', z', t') \\ E'_{y'}(x', y', z', t') \\ E'_{z'}(x', y', z', t') \end{pmatrix} , \quad \boldsymbol{B}'(\boldsymbol{r}', t') = \begin{pmatrix} B'_{x'}(x', y', z', t') \\ B'_{y'}(x', y', z', t') \\ B'_{z'}(x', y', z', t') \end{pmatrix} .$$

The order of the two steps (Lorentz transformation of coordinates and Lorentz transformation of field vectors) does not matter.

Electromagnetic field of a point charge. As an example, we take the electric field of a point charge that is at rest for Alice (see Fig. 14.6). This field is time-independent for Alice and given by Coulomb's law

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^3} \boldsymbol{r} . \tag{14.10}$$

The red vector in Fig. 14.5 shows this field at a particular event $\mathcal{E}$. Note that the vector points away from the charge.

The replacement prescribed in **Step 1** is

$$x \to \gamma_v \cdot (x' + vt') , \quad y \to y' , \quad z \to z'$$

(the time t does not appear in (14.10)), and this implies that

$$\boldsymbol{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \rightarrow \begin{pmatrix} \gamma_v \cdot (x' + vt') \\ y' \\ z' \end{pmatrix}$$

$$r = \sqrt{x^2 + y^2 + z^2} \rightarrow \tilde{r}' := \sqrt{\gamma_v^2 \cdot (x' + vt')^2 + y'^2 + z'^2} \, .$$

Note that $\tilde{r}'$ is not the length of the vector $\boldsymbol{r}'$, but the length of $\boldsymbol{r}$. That is why we put the tilde on top of the r'.

The charge sits at $x' = -vt'$, $y' = z' = 0$ and moves to the left. We restrict ourselves to $t' = 0$.[9] Bob is only interested in the field at $t' = 0$, because the field does not change with time. It moves as a whole to the left, together with the charge. Then,

$$\boldsymbol{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \rightarrow \tilde{\boldsymbol{r}} = \begin{pmatrix} \gamma_v x' \\ y' \\ z' \end{pmatrix} ,$$

$$r = \sqrt{x^2 + y^2 + z^2} \rightarrow \tilde{r}' := \sqrt{\gamma_v^2 x'^2 + y'^2 + z'^2} \, ,$$

and we get

$$\boldsymbol{E}(\boldsymbol{r}') = \frac{1}{4\pi\epsilon_0} \frac{Q}{\tilde{r}'^3} \begin{pmatrix} \gamma_v x' \\ y' \\ z' \end{pmatrix} . \tag{14.11}$$

The vectors $\boldsymbol{E}(\boldsymbol{r})$ and $\boldsymbol{E}(\boldsymbol{r}') := \boldsymbol{E}(\boldsymbol{r}(\boldsymbol{r}'))$ are identical and both sit at $\mathcal{E}$. Only the coordinates of $\mathcal{E}$ are different for Alice and Bob.

Let us take such an event $\mathcal{E}$ in Alice's coordinate system. The electric field vector at this event is shown in Fig. 14.5 in the left diagram. The effect of Step 1 of the transformation of the field is shown in the right diagram of Fig. 14.5. The electric field vector for Alice (dashed red vector) is moved from the coordinates (x, y, z) to the coordinates (x', y', z') of $\mathcal{E}$. The position vector $\boldsymbol{r}$ is now $\tilde{\boldsymbol{r}}' = (\gamma_v x', y', z')$ and still has the same direction and the same length as for Alice. Due to the fact that the factor γ_v only appears with the x'-component, this electric field vector does not lie on a (straight) line from the location of the point charge. The straight line is given by the vector $\boldsymbol{r}' = (x', y', z')$, while the electric field vector is parallel to $\tilde{\boldsymbol{r}}' = (\gamma_v x', y', z')$. Therefore, the field $\boldsymbol{E}(\boldsymbol{r}')$ is not a valid electric field, it is not a solution of Maxwell's equations. But that's not a problem, as Step 2 of the transformation is still missing.

Note again that the Lorentz transformation of the coordinates does not move the electric field vectors. They stay at the same event and only the coordinates change. These coordinates, however, are determined with rods and clocks, and this is what the observer sees. Therefore, the distance from the charge to the event $\mathcal{E}$ is different for different observers.

[9] This means that Bob is interested in events $\mathcal{E}$ with $t' = 0$! Such events, in general, do not lie on Alice's $t = 0$ plane, but they have the t-coordinate $t = \gamma(t' + (v/c^2)x') = \gamma(v/c^2)x' = (v/c^2)x$.

Step 2, the Lorentz transformation of the field vectors according to (14.1) then gives us (note that $\boldsymbol{B} = 0$ for Alice)

$$\begin{aligned}
E'_x(x', y', z') &= E_x(x', y', z') = \frac{1}{4\pi\epsilon_0}\frac{Q}{\tilde{r}'^3}\gamma_v x' , \\
E'_y(x', y', z') &= \gamma_v E_y(x', y', z') = \gamma_v \frac{1}{4\pi\epsilon_0}\frac{Q}{\tilde{r}'^3} y' , \\
E'_z(x', y', z') &= \gamma_v E_z(x', y', z') = \gamma_v \frac{1}{4\pi\epsilon_0}\frac{Q}{\tilde{r}'^3} z' ,
\end{aligned}$$

and we see that the other two components (y' and z') now also get a factor γ_v and $\boldsymbol{E}'(\boldsymbol{r}')$ is eventually proportional to the position vector $\boldsymbol{r}'$ (and therefore lies on a (straight) line through the point charge):

$$\begin{aligned}
\boldsymbol{E}'(\boldsymbol{r}') &= \gamma_v \frac{1}{4\pi\epsilon_0}\frac{Q}{\tilde{r}'^3}\boldsymbol{r}' \\
&= \gamma_v \frac{1}{4\pi\epsilon_0}\frac{Q}{(\gamma_v^2 x'^2 + y'^2 + z'^2)^{3/2}}\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} .
\end{aligned} \tag{14.12}$$

This field is shown in Fig. 14.6 on the right side and is the electric field caused by a uniformly moving electric point charge.

It is instructive to calculate the **curves of constant magnitude** of the electric field (14.12). To do so, we choose $y' = 0$ (which is not a restriction to generality, because the field has rotation symmetry around the x'-axis). With $z' = r' \sin\vartheta$ (see Fig. 14.6 right), we get

$$\begin{aligned}
\gamma_v^2 x'^2 + z'^2 &= \gamma_v^2 r'^2 \left(\frac{x'^2}{r'^2} + \frac{z'^2}{\gamma_v^2 r'^2}\right) \\
&= \gamma_v^2 r'^2 \left(\frac{x'^2}{r'^2} + \frac{z'^2}{r'^2} + \left(\frac{1}{\gamma^2} - 1\right)\frac{z'^2}{r'^2}\right) \\
&= \gamma_v^2 r'^2 \left(1 - \beta^2 \sin^2\vartheta\right) .
\end{aligned}$$

The electric field then reads as

$$\begin{aligned}
\boldsymbol{E}'(\boldsymbol{r}') &= \gamma_v \frac{Q}{4\pi\epsilon_0}\frac{1}{(\gamma_v^2 x'^2 + y'^2 + z'^2)^{3/2}}\boldsymbol{r}' \\
&= \gamma_v \frac{Q}{4\pi\epsilon_0}\frac{1}{\gamma_v^3 r'^3 \left(1 - \beta^2 \sin^2\vartheta\right)^{3/2}}\boldsymbol{r}' \\
&= \frac{Q}{4\pi\epsilon_0}\frac{1 - \beta^2}{\left(1 - \beta^2 \sin^2\vartheta\right)^{3/2}}\frac{\boldsymbol{r}'}{r'^3} .
\end{aligned}$$

The magnitude is then

$$\left|\boldsymbol{E}'(\boldsymbol{r}')\right| = \frac{Q}{4\pi\epsilon_0}\frac{1-\beta^2}{\left(1-\beta^2\sin^2\vartheta\right)^{3/2}}\frac{1}{r'^2}\ .$$

In total, in comparison to the electric field of a resting charge (14.10), the additional factor $(1-\beta^2)/(1-\beta^2\sin^2\vartheta)^{3/2}$ appears. For the special case $\vartheta = 0$, one gets $1/\gamma_v^2 \leq 1$, and for $\vartheta = \pi/2$, one gets $\gamma_v \geq 1$. Therefore, the electric field of a moving charge in comparison to a charge at rest in the direction of motion becomes smaller, while, in the direction orthogonal to it, it becomes larger.

A **magnetic field** also appears now. According to (14.1), it is given by

$$\begin{aligned} B'_x &= B_x = 0\ ,\\ B'_y &= \gamma_v(B_y + \frac{v}{c^2}E_z) = \gamma_v\frac{v}{c^2}E_z\ ,\\ B'_z &= \gamma_v(B_z - \frac{v}{c^2}E_y) = -\gamma_v\frac{v}{c^2}E_y\ , \end{aligned}$$

or, with (14.11),

$$\boldsymbol{B}'(\boldsymbol{r}') = \gamma_v\frac{v}{c^2}\begin{pmatrix}0\\ E_z(\boldsymbol{r}')\\ -E_y(\boldsymbol{r}')\end{pmatrix} = \gamma_v\frac{1}{4\pi\epsilon_0}\frac{Q}{\tilde{r}'^3}\frac{v}{c^2}\begin{pmatrix}0\\ z'\\ -y'\end{pmatrix}\ .$$

Scalar multiplication of $\boldsymbol{B}'$ with $\boldsymbol{v} = (v, 0, 0)$ and with $\boldsymbol{E}'(\boldsymbol{r}')$ shows that $\boldsymbol{B}'$ is perpendicular to the velocity (or the x'-axis) and to the electric field $\boldsymbol{E}'$. In Fig. 14.6 on the right side, the green vectors show the magnetic field on the line $(x, 1, 1)$. The magnetic field has rotational symmetry around the x-axis, similar to the case shown in Fig. 14.3.

Chapter 15
Towards General Relativity

15.1 The Need for a More General Theory

The natural next step now would be to make *Newton's theory of gravitation* compatible with special relativity. Contrary to Maxwell's electrodynamics, Newton's theory is not already in line with special relativity. The reason is the "action-at-a-distance" in Newton's theory. If the Sun were to vanish, according to Newton's theory, the Earth immediately would cease to follow its orbit around the Sun and move along a (straight) line. In special relativity, however, the information that the Sun is not there anymore would travel with the speed of light to the Earth and take a bit more than 8 min to cover this distance. Until the information arrives at the location of the Earth, it would continue following its movement along the orbit around the no longer existing Sun.

We have seen that Newton's mechanics was also inadequate for large velocities and we successfully twirked it to become compatible with special relativity. This, however, is not possible with Newton's gravity. Many attempts have been made and all of them failed.

Einstein had further reasons to look for a more general theory.

We have seen that, according to the relativity principle, physics has to be the same in all inertial systems. The equations of physics must be invariant under Lorentz transformations. If we change from an inertial system to an accelerated reference frame, however, fictitious forces appear (see Sect. 3.3.1). Einstein did not like that inertial systems were preferred reference frames. In his opinion, all reference frames should be equal and the equations of physics should look the same.

And there was a further observation by Einstein that eventually led him towards his theory of gravitation: the fact that the mass plays a double role in physics.

First, the mass of an object appears in Newton's force law: A force $\boldsymbol{F}$ acting on an object causes an acceleration $\boldsymbol{a} = \boldsymbol{F}/m$ in it. We call this the **inertial mass** m_{I}. It is the resistance with which an object opposes the force. For the same force, the larger the inertial mass, the smaller the acceleration.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_15

Second, in a **gravitational field** $\boldsymbol{g}(\boldsymbol{r})$, an object is exposed to a force, the **gravitational force** $\boldsymbol{F}_\mathrm{G}$. This force is proportional to the **gravitational mass** m_G of the object: $\boldsymbol{F}_\mathrm{G} = m_\mathrm{G}\boldsymbol{g}$.

Suppose now that an object is released from rest in a constant gravitational field and starts to fall freely. By "fall freely", we refer to the fact that only gravitation, but no other forces, acts on the object. The object's trajectory is $z(t) = z_0 - \frac{1}{2}at^2$, where a is its acceleration. This acceleration is given by combining the two equations above, which yields

$$\boldsymbol{a} = \frac{m_\mathrm{G}}{m_\mathrm{I}}\boldsymbol{g}.$$

A plethora of experiments, starting with Galilei's Pisa experiments and ranging to contemporary examples, with increasing precision, have shown that, independent of the type of object, its inertial and the gravitational mass are always the same:

$$m_\mathrm{G} = m_\mathrm{I}. \tag{15.1}$$

This implies $\boldsymbol{a} = \boldsymbol{g}$, or that, subject only to the gravitational force, all objects fall in the same way. If, in the vacuum close to the Moon's surface, we release a steel ball and a feather from the same height and at the same time, both arrive at the surface at the same time.

For Einstein, this was too much of a coincidence and required an explanation.

Interestingly, with his **general theory of relativity** (short: **general relativity**, GR), Einstein reached all these objectives: it is a theory of gravitation, it does not single out special reference frames and it explains why the inertial and the gravitational mass of an object are equal.

15.2 Recap of Newton's Theory of Gravitation

Newton's theory of gravitation says that two point masses m_1 and m_2 at a distance r, attract each other by a force along the line intersecting the two point masses and with a strength of

$$F_G = G\frac{m_1 m_2}{r^2}.$$

The force is proportional to each mass and decreases by an inverse-square law with the distance. G is *Newton's gravitational constant.*

If one of the two masses M (e. g., the Earth) is much larger than the other one m (e. g., a satellite), one usually uses a picture in which M creates a **gravitational field** $\boldsymbol{g}(\boldsymbol{r})$. If M is located at the origin of the coordinate system, the gravitational field is given by

$$\boldsymbol{g}(\boldsymbol{r}) = -G\frac{M}{r^2}\boldsymbol{e}_r,$$

where $\boldsymbol{e}_r$ is the unit vector that points from M to the point given by $\boldsymbol{r}$. The other mass m in the gravitational field at $\boldsymbol{r}$ then experiences a force given by

$$\boldsymbol{F}_G(\boldsymbol{r}) = m\boldsymbol{g}(\boldsymbol{r}).$$

The gravitational field in Newton's theory has the property that one can introduce a **gravitational potential** $\Phi(\boldsymbol{r})$ such that

$$\boldsymbol{g}(\boldsymbol{r}) = -\nabla\Phi(\boldsymbol{r}), \tag{15.2}$$

i.e., the gravitational field points in the direction of steepest descent of $\Phi(\boldsymbol{r})$. For the point mass M, the gravitational potential is given by

$$\Phi(\boldsymbol{r}) = -G\frac{M}{r}. \tag{15.3}$$

We can always add a constant to the gravitational potential $\Phi(\boldsymbol{r})$; only differences of this quantity are physically meaningful. Often, this constant is chosen such that the gravitational potential far away from masses becomes zero [as we do in (15.3)].

The gravitational field caused by the Earth on the Earth's surface is given by

$$g := |g(R_E)| = G\frac{M_E}{R_E^2} \approx 10\,\mathrm{m/s}^2,$$

with M_E being the Earth's mass and R_E its radius. The gravitational field points vertically downwards.[1]

The gravitational potential at a height z over the Earth's surface, and normalized such that it vanishes at $z = 0$, is

$$\Phi(z) = -GM_E\left(\frac{1}{R_E + z} - \frac{1}{R_E}\right) \approx G\frac{M_E}{R_E^2}z = gz,$$

and the approximation is valid as long as $z \ll R_E$.

With a theorem from vector calculus (the divergence theorem or Gauss's theorem), Newton's law of gravitation can be reformulated such that it is valid for gravitational fields created not by point masses but by mass distributions and described by *mass density fields*. This alternative formulation is usually called Gauss's law for gravity, but we prefer the term *Newton's field equation* for our purpose. Here it is:

$$\Delta\Phi(\boldsymbol{r}) = 4\pi G\rho(\boldsymbol{r}). \tag{15.4}$$

[1] Actually, the direction of the field defines the meaning of "vertically downwards".

The operator Δ is the Laplace operator, which performs the second derivative in space. Newton's field equation yields the gravitational potential $\Phi(\boldsymbol{r})$ for a given mass density $\rho(\boldsymbol{r})$.

15.3 The Equivalence Principle

15.3.1 The Equivalence Principle

Einstein referred to it as the happiest thought of his life, and indeed it served him as a guide on his way to general relativity.

His thought was based on the experimentally confirmed equality of inertial and gravitational mass. In a free falling cabin, an observer will not be able to detect a gravitational effect on the movement of objects in the cabin because the cabin and all objects in it fall in the same way. In other words: for the observer in the cabin, there is no gravitational field. It is "transformed away" by the acceleration caused by the gravitational field. The equivalence of inertial and gravitational mass tells us that this is true for the movement of objects, but it does not make a statement about other possible influences of the gravitational field on the physics in the cabin.

And here, Einstein generalized the observation to all physical effects. In his **equivalence principle** (EP),[2] he stipulated that, in general, the physics in an accelerated cabin that freely falls in a gravitational field is the same as physics in an inertial frame without gravity. This implies that we can use what we learned in the preceding sections of this book to describe physics in a gravitational field but have to replace the inertial system with a freely falling system.

The acceleration of an object in a large accelerated cabin is independent of the position of the object in the cabin. Therefore, the "acceleration field" is homogeneous. A gravitational field, however, is never constant in a larger region of space. Therefore, the accelerated cabin can compensate the gravitational field only locally. That is why the cabin should be small and why the reference frame of the cabin provides a **local inertial frame** (LIF).

Figure 15.1 illustrates this. The objects A, B, and C in a large cabin E fall freely in the gravitational field of the Earth. The cabin falls such that B is exactly at rest in it. The objects, however, fall in slightly different directions (left figure), which results in A and C approaching B in the cabin frame. The forces that make A and C approach B in the cabin frame are called **tidal forces**. A local inertial frame is sufficiently small enough to render the tidal forces negligible.

[2] Sometimes, $m_{\mathrm{G}} = m_{\mathrm{I}}$ is referred to as the *weak* equivalence principle, while the generalization to all physics is called the *strong* equivalence principle.

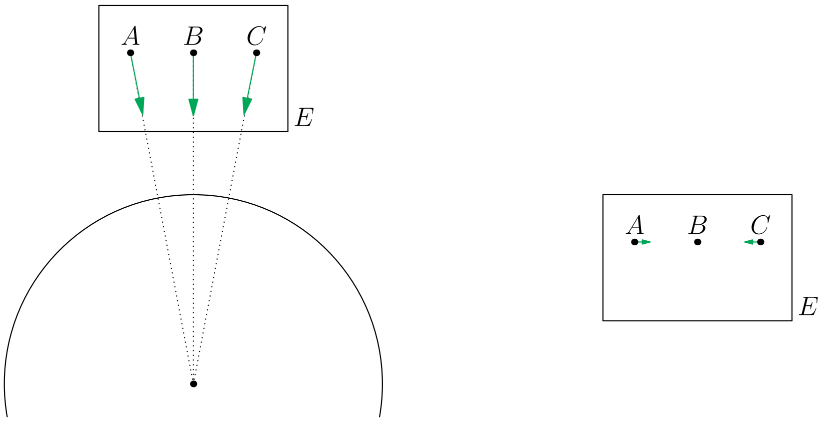


Fig. 15.1 Tidal forces and the extent of local inertial frames. Left: cabin E and objects A, B, C fall freely. Right: in the rest frame of the cabin E, where B is at rest, the objects A and C move toward B

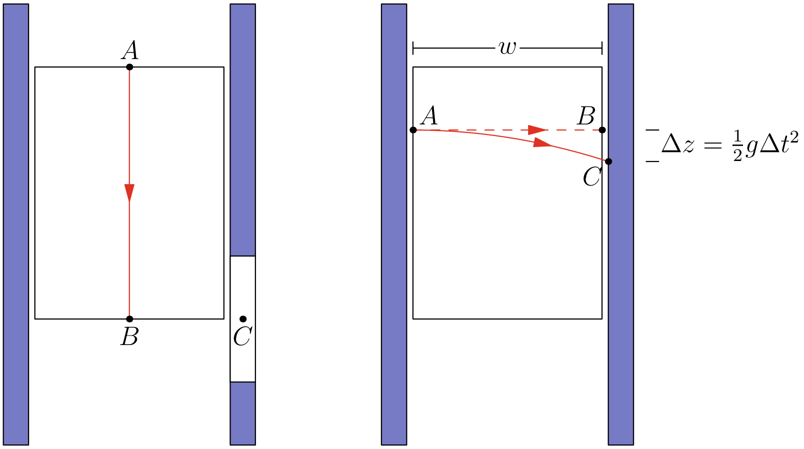


Fig. 15.2 Free-falling elevator in the elevator shaft. Left: with a light ray sent vertically from A to B. Right: with a light ray sent horizontally from A to B

15.3.2 Consequences from the Equivalence Principle

Gravitational frequency shift. Suppose at time $t = 0$, an elevator cabin is released from rest and starts falling freely in its shaft (see Fig. 15.2, left). At the same time, a source A at the ceiling of the elevator sends a light ray with frequency ω_{A} vertically downwards. Let h be the height of the elevator cabin; then, the light ray will arrive approximately at time $t_{\mathrm{B}} = h/c$ at the detector B, which sits on the floor of the cabin. Due to the fact that the cabin forms a local inertial frame, the detector measures the frequency $\omega_{\mathrm{B}} = \omega_{\mathrm{A}}$, i. e., the same frequency as emitted by A.

Incidentally, at t_B, the cabin floor passes by a door in the elevator shaft, where a further detector C is placed. At t_B, the cabin has the velocity $v = gt_\text{B} = gh/c$ relative to the elevator shaft and C. Therefore, in the local inertial frame, C moves with this velocity relative to detector B and the light ray and, according to the Doppler effect, the detector C measures the higher frequency[3]

$$\omega_\text{C} \approx \omega_\text{A} \cdot \left(1 + \frac{v}{c}\right).$$

With the frequency shift $\Delta\omega := \omega_\text{C} - \omega_\text{A}$ and the potential difference $\Delta\Phi = \Phi_\text{C} - \Phi_\text{A} = -gh$, we arrive at

$$\frac{\Delta\omega}{\omega_\text{A}} = \frac{v}{c} = \frac{gh}{c^2} = -\frac{\Delta\Phi}{c^2}.$$

This is the **gravitational frequency shift**. It is proportional to the potential difference. Note that this agrees with (9.15) in our discussion of the influence of gravitation in the experiment by Hafele and Keating.

When the source A sent the light ray, it was at rest relative to C. Therefore, we can conclude that a light wave, which travels along the direction of the gravitational field, downwards to lower gravitational potentials, increases its frequency. It becomes *blueshifted*.[4] We can reverse our Gedanken experiment with the falling elevator and will find out that a light wave that travels along the direction of the gravitational field upwards to higher gravitation potentials, decreases its frequency. It becomes *redshifted*.

For the "usual" gravitational fields, the effect is very tiny. A light ray sent from the Earth's surface to a place in the Universe far from matter experiences a redshift of $\Delta\omega/\omega = -\Delta\Phi/c^2 = (GM_E/R_E)/c^2 \approx 10^{-9}$.

The experiment of Pound and Rebka. There is a famous experiment, carried out in 1959 by the US-physicist Robert Pound and his graduate student Glen Rebka, in which the gravitational frequency shift was demonstrated for the first time.

The basic idea of the **Pound-Rebka experiment** is simple. Inside of a tower, electromagnetic radiation (photons) of a particular frequency was emitted from a source at the top of the tower and sent to the bottom of the tower, where it was received by a detector that was located 22.5 m below the emitter. As discussed above, general relativity predicts a tiny shift toward a higher frequency (blueshift) of

$$\frac{\Delta\omega}{\omega} = \frac{gh}{c^2} = 2.5 \times 10^{-15}.$$

[3] We only make an approximate calculation here, therefore it is fine to use the classical formula for the Doppler effect.

[4] Blue light has a higher frequency than red light. Therefore: *blueshifted* (*redshifted*) refers to a shift toward higher (lower) frequencies.

To be able to detect such a small shift, photons of high frequency are in order. Pound and Rebka used gamma photons with an energy of 14.4 keV. Visible light has an energy of about 1.5 eV (red) to 3 eV (blue), so the gamma photons had a frequency of about 10.000 times larger than that of red light.

In the domain of visible light, the photons emitted by an atom via a particular state transition from an excited state to the ground state can be detected by the same type of atom, which, if the frequency of the photons is correct, is excited from the ground state and scatters the photons intensely in all directions. This is called resonance fluorescence. But if the emitter is not at the same gravitational potential as the detector, the photons from the emitter will no longer be in exact resonance with the detector. In order to be able to continue using the detector, it will need to be moved. Then, depending on the detector's velocity, the Doppler effect will compensate for the frequency shift of the photons.

For high-energy radiation, this method does not work as is. The reason is that when an atom emits a high-energy photon, the atom becomes recoiled into a more or less random direction. And this recoil, again by the Doppler effect, changes the frequency of the emitted photon by a small random value. This leads to the fact that, even when the detector is off in frequency, it still will detect photons, making the experiment useless.

To the rescue came a discovery, made just two years prior to Pound's and Rebka's experiment by the German physicist Rudolf Mössbauer, who discovered how this recoil effect can be rendered harmless: by cooling the crystal, formed by the emitting atoms, down to very low temperatures. The explanation is that, in this case, the emitting atom can no longer move independently of the crystal; only the crystal as a whole can move. For that reason, the recoil is absorbed not just by the atom, but by the much more massive crystal. This leads to neglectably low recoil velocities and the frequencies of the emitted photons do not spread over a broad band.

To conclude, in Pound's and Rebka's experiment, on the top of the tower, an emitter at very low temperature emits photons of an energy of 14.4 keV that travel down the tower and experience a blueshift. The detector, which is also cooled, is moved away from the photon. The velocity of the detector is changed until the absorption is maximal. Then, the detector's velocity, via the Doppler effect, exactly compensates the gravitational blueshift of the photons. Pound and Rebka found the optimal detector velocity to have the very small value of 7.5×10^{-7} m/s, in accordance with the predictions of general relativity.

Gravitational time dilation. In Sect. 9.4, we saw that an atomic clock is nothing but a device that counts the periods of oscillations of certain atoms. Suppose that our clock can also emit electromagnetic radiation with this frequency, which would be microwave radiation.[5] Suppose further that our clock can also measure the frequency of received microwave radiation.

[5] We could, in principle, also use visible light, but counting the very fast oscillations of an atom emitting light is very difficult.

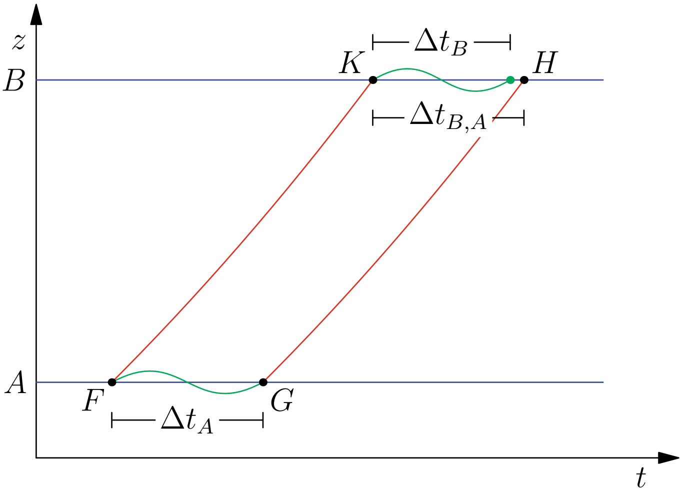


Fig. 15.3 Gravitational time dilation

Take a location A close to the Earth's surface and another one B vertically above it such that there is a height difference h (see Fig. 15.3). Both locations are fixed relative to the Earth, their mutual distance does not change and there is no time dilation from special relativity.

Suppose that, at A, we have two clocks of the same type. Then, we move one clock up to location B. Let the periods of the respective clocks (measured next to the clock) be Δt_A and Δt_B. Due to the fact that the clocks are equal, we have $\Delta t_A = \Delta t_B$.

Now, clock A sends microwaves to clock B. The world lines of two subsequent wave nodes (with phase $\varphi = 0$) are drawn in Fig. 15.3. If there was no gravitational field, the world lines would be (straight) lines at 45° to the t- and z-axes. But there is a gravitational field, which somehow will influence the radiation's world line. We have drawn the lines with a somewhat arbitrary form, as we do not know this influence yet. What we do know, however, is that the gravitational field is static (independent of time), and therefore both world lines must be "parallel".

The frequency of the microwave emitted by clock A at A is ω_A. On its way to location B, this microwave experiences a redshift, its frequency $\omega_{A,B}$, measured by clock B, is a bit smaller than ω_A. For one wave period, we have $\omega \Delta t = 2\pi$. Therefore, the period $\Delta t_{A,B}$ of the wave emitted by A and measured by B is

$$\Delta t_{A,B} = \frac{\omega_A}{\omega_{A,B}} \Delta t_A = \left(1 + \frac{\Phi_A - \Phi_B}{c^2}\right)^{-1} \Delta t_B > \Delta t_B .$$

For B, the clock at A runs more slowly.

But is this really the case? Imagine you are at B and see the light coming from the clock at A redshifted. Then, you could argue that clock A runs at the same pace as clock B but, due to the redshift, the microwave carries the wrong information to you. However, this argument can easily be refuted. Suppose both clocks are synchronized at B and you bring one clock down to location A. This transport can be done in a way that it does not influence the clock's time by much. Then, you wait a long time. While your clock B ticks many times, you receive fewer ticks from the other clock at A (due to the redshift). Now, you bring the clock back to B, again with only a small influence on its time. You will notice that, on clock B, less time will have passed than on the clock that has been left at A. Therefore, gravitational time dilation is real. The clock at A runs more slowly than the one at B.

Gravitational frequency shift implies curved spacetime. We have seen already that the idea that clocks at higher gravitational potential go faster is real, and it has been shown in the experiment of Hafele and Keating. Without taking gravitational time dilation into account, satellite navigation would not work.

Let us again have a look at the spacetime diagram in Fig. 15.3. We see that the figure $F - G - H - K$ forms a kind of a parallelogram, and therefore the distances $\overline{FG}$ and $\overline{KH}$ should be equal. This is not true, however, because, from event K to H, *more time* passes (for the observer at B) than from event F to G (for observer A). Therefore, spacetime must necessarily be curved.

Hence: the gravitational frequency shift (which has been observed in the experiment of Pound and Rebka and many others and is also essential for the functioning of GPS) *implies that spacetime is curved.*

Local bending of light. Let us go back to the free-falling elevator (see Fig. 15.2, right). This time, in the elevator, Alice sends a light ray from A to B, horizontally through the elevator. The light travels on a (straight) line because the elevator forms a local inertial frame.

Bob, who is at rest relative to the elevator shaft, sees something different. The light ray leaves location A and, due to the fact that the light ray needs the finite time $\Delta T = w/c$, where w is the distance from A to B, it arrives at location C. For Bob, the light ray follows a parabola

$$z(x) = \frac{g}{2c^2}x^2,$$

which has the curvature $\kappa = g/(2c^2)$ at $x = 0$ (Fig. 15.4).

Bending of light passing by the Sun and gravitational lenses. Two consequences of this bending of light rays in the gravitational field were among the first confirmations of general relativity.

One is the bending of light that passes by large massive objects in the universe. Einstein, in 1915, had predicted that a light ray that passes by the Sun, almost grazing its surface, would be deflected by an angle of about $1.75''$.[6] This is a very small angle,

[6] An initial calculation by Einstein, made in 1911 and considering only the equivalence principle as we did above, predicted a wrong value of half of the correct one. The reason is that both the curvature

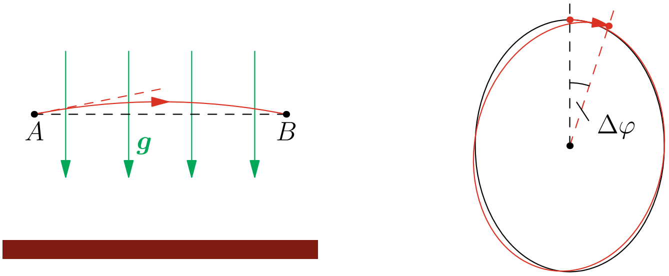


Fig. 15.4 Trajectory of a light ray from A to B in a uniform gravitational field pointing in the negative z-direction. Left: the light ray follows a geodesic and is bent in the direction of the gravitational field. Right: in Newton's gravity, a light mass point orbiting a very heavy central mass point draws an ellipse around the central mass. In Einstein's gravity, the very small deflection of light causes the axes of the ellipse to rotate very slowly

but the British astronomer Arthur Eddington was able to confirm this in a famous experiment conducted during a total solar eclipse in 1919.

Heavier masses like galaxies, or clusters of galaxies, have a much larger influence on the trajectory of light, which resembles that of optical lenses. **Gravitational lenses** were first observed in 1979.

Anomalous perihelion precession of Mercury. There is another effect of spacetime curvature that is directly observable in the solar system and that was also an early confirmation of general relativity.

According to Newton's gravity, a planet moves on an ellipse, with the Sun being located in one of the foci. The major axis of the ellipse connects to points on the planet's orbit, the perihelion, which is the point closest to the Sun, and the aphelion, which is the most distant one. According to Newton's theory, this major axis is fixed in space. Observations made as early as in the 19th century, however, showed that the perihelion actually rotates. This rotation is called **Mercury's perihelion precession**. A large part of this precession could be attributed to the effect of other planets on Mercury's orbit, but there was a remainder of about $43''$ per century left. In 1915, with his new theory, Einstein calculated the effect of curved spacetime on Mercury's orbit and got the experimentally determined value.

15.4 Curved Surfaces

The basis of general relativity is curved (or warped) four-dimensional spacetime. This is difficult to imagine, because our imagination ends with three dimensions, and the

in time and the curvature in space contribute to the bending in the same way. The equivalence principle is able to capture the curvature in time, but not the one in space. To do so, Einstein's full theory of gravitation is necessary.

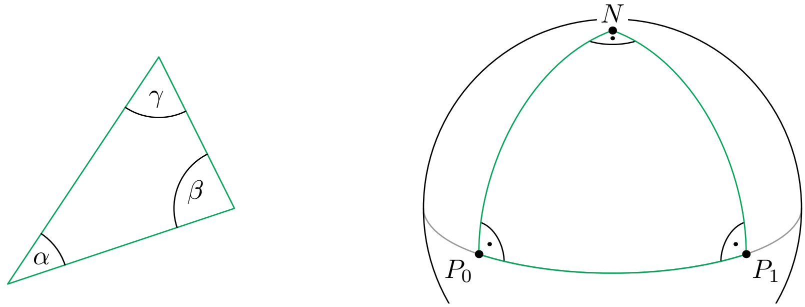


Fig. 15.5 A triangle on the plane and another one on the sphere

fact that spacetime is curved makes this even worse. Fortunately, many of the concepts used in general relativity can be learned simply by studying two-dimensional surfaces. The main contributor to the mathematics of two-dimensional surfaces was Carl Friedrich Gauss. Gauss's student Bernhard Riemann later generalized *differential geometry* to arbitrary dimensions.

15.4.1 The Geometry of Curved Surfaces

In classical physics (including special relativity), it is assumed that **space is flat** (and geometry is *Euclidean*). This means, e. g., that the angle sum of any triangle is 180° (or π). In particular, on a flat *surface* (a plane), the angle sum of a triangle is 180°.

This is not the case for general curved surfaces. Consider creatures that live on a sphere and that only experience the two dimensions of the sphere. We humans also live on the surface of a sphere (the surface of the Earth), but this surface is "embedded" in three-dimensional space, and we can leave the surface "upward" with a ladder, a plane or a rocket or "downward" with a spade. Our creatures cannot do this.

Suppose that the creatures draw a *triangle*, three (straight) lines on the surface. The obvious first question that arises is: what is a "(straight) line" on a curved surface? The generalization of a (straight) line between two points on a flat surface to a curved surface is the shortest curve on the curved surface between these two points. Such curves are called *geodesics*. If one of our creatures walks on the surface and always "follows its nose", it automatically treads a geodesic.

For the triangle (see Fig. 15.5), we draw one (straight) line from the North pole N to point P_0 at 0° longitude on the equator, the second again from the North pole to point P_1 at 90° longitude on the equator, and eventually the third on the equator from P_0 to P_1. The three lines form a triangle, and at each of the three points, the two lines enclose an angle of 90° ($\pi/2$). Therefore, the sum of the angles of this triangle is 270° ($3\pi/2$), and not 180° as in flat space. Geometry on the sphere is different from Euclidean geometry on the plane!

Carl Friedrich Gauss, who was probably the most important mathematician of the modern age, has shown that the geometry on a two-dimensional surface is completely characterized by specifying the **curvature** at each point P on the surface. This curvature can be determined by drawing a (small) triangle around the point P, determining its *area* σ and the *excess angle* $\epsilon := \alpha + \beta + \gamma - \pi$, where α, β, γ are the triangle's angles, and making the triangle smaller and smaller. In the limit of a vanishingly small triangle, the curvature is given by $\kappa := \epsilon/\sigma$ and it is completely independent of the type of triangles that you use to determine κ.

On the sphere, the curvature κ is the same at each point. To determine κ, we can therefore use the large triangle constructed by our creatures above with the excess angle $\epsilon = 3 \cdot \pi/2 - \pi = \pi/2$ and the surface area $\sigma = \pi R^2/2$ (R is the radius of the sphere; the surface of the triangle is $1/8$ of the surface of the sphere), which gives us $\kappa = 1/R^2$. As $R > 0$, the curvature of a sphere is positive. And, as intuitively expected, the smaller the sphere, the larger the curvature. The Earth is huge, and this is why, in all but the most minute way, it seems flat to us.

We humans experience the Earth's surface as a two-dimensional surface that is *embedded* in three-dimensional space, and we have another method to determine the curvature of a surface. Take a general two-dimensional surface (see Fig. 15.6), with a point P on it. Then, there is a well-defined plane, which is tangent to the surface at P, the *tangent plane*, and a vector, the *surface normal*, which is normal to the plane in P. Consider a *normal plane*, which is a plane that contains the surface normal. The intersection of this plane with the surface is a curve through P. We can determine the curvature of this curve at P, which is the radius of the circle, which fits the curve at P. We can do the same for different normal planes (all of which include the surface normal) and will find out that there is a maximal curvature r_1 and a minimal curvature r_2. These are the *principal curvatures* of the surface at P. The related normal planes are mutually perpendicular and are called the *planes of principal curvature*. In Fig. 15.6, we have $r_1 > 0$ and $r_2 < 0$.

We see that, if we consider the surface embedded in three-dimensional space, we need *two* curvatures r_1 and r_2 to describe it. Our creatures, however, need only one curvature κ. Gauss has shown, in his *theorema egregium*, that the *Gaussian curvature* κ is given by $\kappa = 1/(r_1 r_2)$ and that it completely describes the geometry on the surface. The individual principal curvatures are not needed to describe the geometry on the surface. In the case of the sphere, we have $r_1 = r_2 = R$, and therefore $\kappa = 1/R^2$.

Why are the individual radii not needed for the description of the surface? Take a plane surface. You can draw a triangle and notice that the angle sum is π, and therefore, as expected, the curvature vanishes. If you roll the sheet into a cylinder surface, the edges of the triangle are still (straight) lines (geodesics), and the angles at the triangle's vertices do not change. Therefore, the Gaussian curvature κ stays the same. Triangles on a cylinder have an angle sum of π and the cylinder's curvature vanishes. This is consistent with Gauss' formula $\kappa = 1/(r_1 r_2)$ connecting the principal curvatures with the Gaussian curvature. In the case of the cylinder surface, one of the two planes of principal curvature through a point P on the surface is perpendicular to the axis of the cylinder and the other one contains this axis. The curve

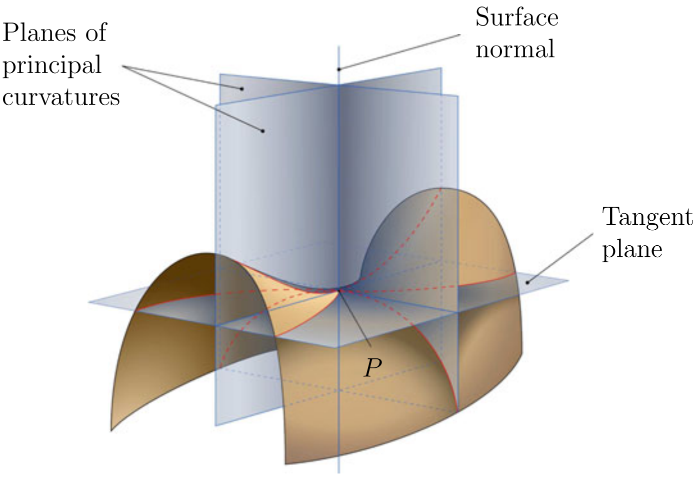


Fig. 15.6 A curved surface with the tangent plane, the surface normal and the two planes of principal curvature at P

given by the intersection of the cylinder surface and the plane containing the cylinder axis is a line and has no curvature. Therefore, the maximal curvature r_1 is infinitely large. The curve given by the intersection of the plane perpendicular to the cylinder axis and the cylinder surface is a circle and the minimal curvature r_2 corresponds to the radius of the cylinder. In the end, the Gaussian curvature becomes zero.

The Gaussian curvature is also called the *internal curvature* of the surface, and this is indeed the only curvature that can be determined by our two-dimensional creatures. *Locally*, they see no difference between a flat surface (plane) and the cylinder surface. *Globally*, however, there is a difference. If the creatures start at a point P on the cylinder surface and walk in an appropriate direction, after some time, they will return to point P. This is not possible on a plane. The two curvatures r_1 and r_2 are *external curvatures* that only exist if the surface is embedded in a three-dimensional space, and this external curvature even locally distinguishes between the plane and the cylinder surface.

Besides the triangles, *parallel lines* or *circles* are other practical indicators of a curved surface (see Fig. 15.7). On planes, parallel lines never meet. The situation is different for curved surfaces. All the meridian lines in the geographic coordinate system (the lines connecting the North pole with the South pole) on the Earth's surface are parallel at the equator. Nevertheless, these parallel lines meet at the North and at the South pole.

Instead of a triangle, you can also draw a *circle* with a certain radius on the surface. This also allows the creatures to determine the Gaussian curvature. Take a *circle of latitude* on the Earth's surface, which is a circle around the North pole. The circumference of such a circle of latitude with latitude φ is

$$S(\varphi) = 2\pi R_\text{E} \sin\varphi,$$

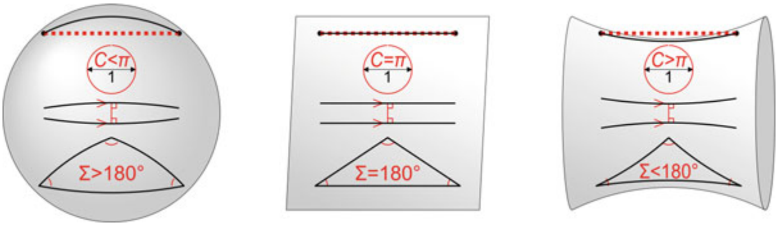


Fig. 15.7 Surfaces of different curvature and indicators for the curvature. Left: surface with positive curvature; Middle: flat surface; Right: surface with negative curvature

where R_E is the radius of the Earth. Additionally, the radius of the circle is equal to the distance from the North pole to the circle of latitude (as measured on the Earth's surface) and is $r(\varphi) = \varphi R_\mathrm{E}$, and therefore

$$S(r) = 2\pi R_\mathrm{E} \sin(r/R_\mathrm{E}).$$

Gauss has shown that the curvature κ is given by the excess circumference $2\pi r - S(r)$ divided by r^3, in the limit of vanishing radius r, i. e.,

$$\kappa = \frac{3}{\pi} \lim_{r\to 0} \frac{2\pi r - S(r)}{r^3}.$$

For the considered sphere, this yields (with $x = r/R_E$)

$$\begin{aligned}\kappa &= \frac{3}{\pi} \lim_{r\to 0} \frac{2\pi r - 2\pi R_\mathrm{E} \sin(r/R_\mathrm{E})}{r^3} \\ &= \frac{6}{R_\mathrm{E}^2} \lim_{x\to 0} \frac{x - \sin x}{x^3} = \frac{6}{R_\mathrm{E}^2} \lim_{x\to 0} \frac{x - (x - x^3/6 + \cdots)}{x^3} \\ &= \frac{1}{R_\mathrm{E}^2}.\end{aligned}$$

Flat surfaces yield Euclidean geometry and surfaces with positive curvature, as the sphere discussed above, yield *elliptic geometry*. In contrast, surfaces with negative curvature yield *hyperbolic geometry*. In hyperbolic geometry, triangles have an angle sum of less than π, circles have a circumference larger than 2π times their radius, and there exist lines that are not parallel and, nevertheless, intersect nowhere (see Fig. 15.7).

As the sphere is the surface with constant positive curvature, the *pseudosphere* is the surface with constant negative curvature. The pseudosphere cannot be embedded in three-dimensional space, but the *tractricoid* (which results from revolving a tractrix about its asymptote), shown in Fig. 15.8, is very close to it. Its curvature is constant everywhere, with exception of its equator.

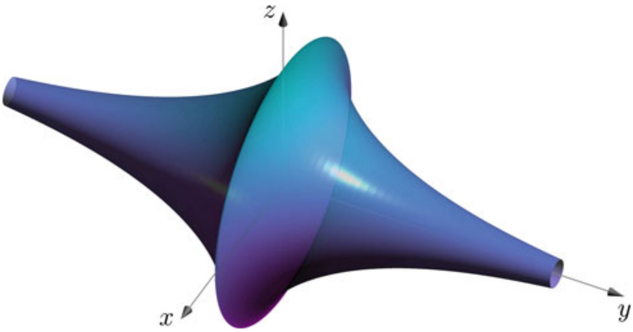
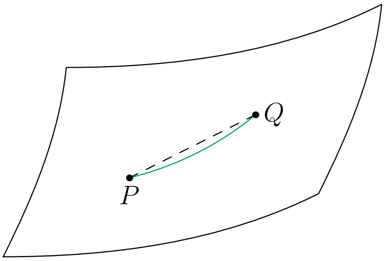


Fig. 15.8 The surface shown is a tractricoid. It is the closest surface to a pseudosphere that is embeddable in three-dimensional space

Fig. 15.9 The distance of P and Q on the curved surface becomes its distance in the three-dimensional embedding space if P and Q are infinitesimally close



15.4.2 *Quantitative Description of Curved Surfaces*

Gauss coordinates. We turn our discussion of curved surfaces towards a more quantitative topic now. To describe a surface, each point of the surface is identified by a coordinate pair (u, v) in such a way that the coordinate pairs lie continuously on the surface. Such coordinates are called **Gauss coordinates**. An example for a plane could be the cartesian coordinates (x, y) or the polar coordinates (r, φ) and an example for the surface of the unit sphere is the spherical coordinates (φ, ϑ) with the azimuthal and polar angles.

A surface[7] is specified by giving the location of each of the surface points in three-dimensional space, i. e.,

$$x(u, v), \quad y(u, v), \quad z(u, v). \tag{15.5}$$

The metric tensor. The distance between two infinitesimally close points $P = (u, v)$ and $Q = (u + du, v + dv)$ on the surface (i. e., the length of the shortest line connecting P and Q on the surface) is then given by the distance of the points in 3D space (see Fig. 15.9). The square ds^2 of this distance ds is

[7] Provided that it can be embedded into three-dimensional space.

$$ds^2 = dx^2 + dy^2 + dz^2.$$

Using the notation $x_u := \partial x/\partial u$ etc. for the partial derivatives of (15.5), one gets

$$dx = x_u\, du + x_v\, dv, \text{ etc.}$$

Therefore, ds^2 can be written as

$$ds^2 = (x_u^2 + y_u^2 + z_u^2)\, du^2 + (x_u x_v + y_u y_v + z_u z_v)\, du\, dv + (x_v^2 + y_v^2 + z_v^2)\, du^2$$

or

$$ds^2 = g_{uu}\, du^2 + 2g_{uv}\, du\, dv + g_{vv}\, dv^2 \tag{15.6}$$

and is sometimes called the *first fundamental form*. The quantities g_{uu}, g_{uv}, g_{vv} and $g_{vu} := g_{uv}$ are the **metric tensor** (in g_{uu} etc., the indices refer to the components of g, not to partial derivatives). *Given the metric tensor, the geometry on the surface is fully specified.* This includes all distances on the surface and all angles. We can write the metric tensor in matrix form

$$\hat{G} = \begin{pmatrix} g_{uu} & g_{uv} \\ g_{vu} & g_{vv} \end{pmatrix} \tag{15.7}$$

and the first fundamental form becomes

$$ds^2 = \begin{pmatrix} du & dv \end{pmatrix} \hat{G} \begin{pmatrix} du \\ dv \end{pmatrix}.$$

As a first example, we use a plane (flat surface). With cartesian coordinates, the first fundamental reads as

$$ds^2 = dx^2 + dy^2,$$

while, in the case of polar coordinates, we have

$$ds^2 = dr^2 + r^2\, d\varphi^2.$$

In both cases, there is no "mixed term" $du\, dv$, i.e., $g_{uv} = g_{vu} = 0$, and the metric tensor is diagonal. This means that the coordinate lines are always mutually perpendicular. If the metric tensor is equal to the unit matrix, we have cartesian coordinates.

In the case of the polar coordinates, the metric tensors depend on the position, although the surface is flat. This indicates that it cannot be read easily from the metric tensor whether or not the surface is flat—although the metric tensor contains this information. For the same surface, different coordinates result in different metric tensors.

As a second example, we take the sphere of (fixed) radius R. Using spherical coordinates (φ, ϑ), the sphere is given by

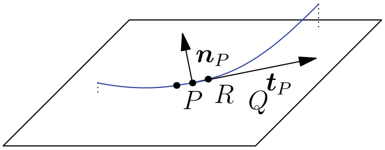


Fig. 15.10 Curve on a surface. The tangent vector $\boldsymbol{t}_P$, the principal normal $\boldsymbol{n}_P$ and the osculating plane to the curve at P are shown

$$\begin{aligned} x(\varphi, \vartheta) &= R \sin \vartheta \cos \varphi, \\ y(\varphi, \vartheta) &= R \sin \vartheta \sin \varphi, \\ z(\varphi, \vartheta) &= R \cos \vartheta. \end{aligned}$$

From this, the first fundamental form

$$ds^2 = R^2(d\vartheta^2 + \sin^2 \vartheta \, d\varphi^2)$$

follows. Close to the equator, we have $\vartheta \approx \pi/2$ and the first fundamental form becomes $ds^2 = R^2(d\vartheta^2 + d\varphi^2)$, which shows that the metric tensor is approximately constant and the coordinates are approximately cartesian.

Curvature. Gauss has shown, in his *theorema egregium*, that the curvature κ is indeed independent of the used coordinates and has given a formula that yields κ as (a rather complicated) function of the first and second partial derivatives of the metric tensor. A change of coordinates changes the metric tensor but leaves the curvature invariant.

For the sphere, Gauss's formula yields $\kappa = 1/R^2$.

Geodesics. Take an arbitrary curve

$$\boldsymbol{r}(s) = (x(s), y(s), z(s))$$

with the curve parameter s chosen such that it measures the length of the curve (see Fig. 15.10). The first derivative $\dot{\boldsymbol{r}}(s)$ at the point P is a unit vector (that's because we chose the curve length as parameter) that is tangent to the curve and called the curve's *tangent vector* $\boldsymbol{t}_P$ at P. The second derivative $\ddot{\boldsymbol{r}}(s)$ is the *curvature vector* $\boldsymbol{k}_P$ at P. It vanishes if the curve is straight in P. Otherwise, it is perpendicular to the tangent vector and its length is equal to the *curvature* κ of the curve at P. The curvature vector is usually written as $\boldsymbol{k}_P = \kappa \boldsymbol{n}_P$, the product of the curvature and the *principal normal* $\boldsymbol{n}_P$. The two vectors $\boldsymbol{t}_P$ and $\boldsymbol{n}_P$ span the *osculating plane* at P, which has the property that, at P, the curve lies within it. If we take two points Q and R on the curve before and after P, but infinitesimally close to it, all three points lie in the osculating plane.

If the curvature vector $\boldsymbol{k}_P$ of the curve at each point P on the curve coincides with the surface normal at P, the curvature of the curve is always perpendicular to the surface and the projection of the curvature vector to the tangent plane vanishes. For the creatures that live on the surface, the curve then is straight. Such a curve is called a **geodesic**. Geodesic curves are the closest to (straight) lines on curved surfaces. On the sphere, the geodesics are the intersections of planes through the origin of the sphere with the sphere (great circles).

Through each point$P$ on a surface, there is exactly one geodesic for each direction in the tangent plane of the surface. A point P on the surface and a tangent vector uniquely determine a geodesic.

We already mentioned that a geodesic can also be characterized by the fact that it is the shortest curve connecting two points on the surface.

Suppose that the shortest curve from A to B is given by

$$(u(\lambda), v(\lambda)),$$

where $\lambda \in [0, 1]$ and $\lambda = 0$ corresponds to A while $\lambda = 1$ corresponds to B.

Then, the length s of this curve $\mathcal{C}$, given by

$$s = \int_{\mathcal{C}} ds = \int_0^1 \sqrt{\left(\frac{du}{d\lambda}\right)^2 + \left(\frac{dv}{d\lambda}\right)^2}\, d\lambda,$$

must be the shortest for all curves from A to B. This is a minimum principle and a standard problem in mathematics. It leads to the Euler-Lagrange differential equation, which, in the case at hand, is

$$\frac{d^2u}{d\lambda^2} + \Gamma^1_{11}\left(\frac{du}{d\lambda}\right) + 2\Gamma^1_{12}\frac{du}{d\lambda}\frac{dv}{d\lambda} + \Gamma^1_{22}\left(\frac{dv}{d\lambda}\right) = 0, \tag{15.8}$$

$$\frac{d^2v}{d\lambda^2} + \Gamma^2_{11}\left(\frac{du}{d\lambda}\right) + 2\Gamma^2_{12}\frac{du}{d\lambda}\frac{dv}{d\lambda} + \Gamma^2_{22}\left(\frac{dv}{d\lambda}\right) = 0. \tag{15.9}$$

These differential equations together are called the *geodesic equation*. Given the metric tensor g_{ij} of a curved surface, the *Christoffel symbols* Γ^k_{ij}, which are a combination of the metric tensor with its derivatives, can be calculated. Then, the geodesics follow from the geodesic equation.

15.5 Curved Spacetime and General Relativity

15.5.1 Curved Surfaces Versus Curved Spacetime

In general relativity, space and time form a four-dimensional space (space here in the mathematical sense) called *spacetime*, which is curved (warped) through the presence of masses and energies. The obvious difference from curved surfaces is the two additional dimensions, which complicate matters a lot. We have seen that, in curved surfaces, one number is sufficient to specify the curvature at a point P. In four-dimensional spacetime, however, 20 numbers are needed. But there is a more important difference in the case of curved surfaces. In our discussion on surfaces, the non-curved case is given by the metrics $ds^2 = dx^2 + dy^2$, which corresponds to Euclidean geometry on the plane. In curved spacetime, the non-curved case is given by special relativity and (in two-dimensional spacetime) the metrics (see 9.3)

$$ds^2 = c^2\,dt^2 - dx^2.$$

This is called the **Minkowski metrics** and it is the metrics for spacetime without a gravitational field or masses and energies. It is the starting point for general relativity.

In general relativity, the trajectory taken by a free-falling object upon which gravity, but no other forces, acts is given by a geodesic in spacetime. For example, the path of an object that we throw upwards and that falls down to Earth must be a geodesic in curved spacetime.

To demonstrate how Einstein's general relativity accomplishes this, we first stay in classical mechanics and discuss a more unusual way of determining trajectories of objects.

15.5.2 The Principle of Stationary Action and Geodesics

Can we construct a non-Euclidean spacetime (x, t) such that the trajectory of a freely falling object is a geodesic? Yes, that's possible, as Einstein has demonstrated. We cannot prove this here, but we can explain the idea. It is possible to knead the *principle of stationary action* into a statement about geodesics.

If, in classical mechanics, you want to determine the trajectory $\boldsymbol{r}(t)$ of a particle that is subject to a certain force $\boldsymbol{F}$, you solve Newton's force law, which reads as $\ddot{\boldsymbol{r}}(t) = \boldsymbol{F}/m$. This gives you two integration constants, which you use to accomodate for the initial conditions, the position $\boldsymbol{r}_0$ and the velocity $\boldsymbol{v}_0$ of the particle at t_0.

There's an alternative way to determine the trajectory of a particle. Here, two points P_0 and P_1 in spacetime are given and one determines the trajectory that passes through these two points.

Let's restrict ourselves to one space dimension, measured by the z-coordinate, and consider the two-dimensional z-t-spacetime (see Fig. 15.11, left). Then, the two points will be $P_0 = (z_0, t_0)$ and $P_1 = (z_1, t_1)$ and we seek for the particle trajectory that goes through these two points.

To make this idea clearer, let's consider the trajectory

$$z(t) = -\frac{1}{2}g(t - t_0)^2 + v_0(t - t_0) + z_0,$$

which is that of a particle falling freely (initially thrown upwards) in a uniform gravitational field $\boldsymbol{g}$ and that has the position z_0 and velocity v_0 at time t_0. At some later fixed time t_1, it is at $z_1 = -g(t_1 - t_0)^2/2 + v_0(t_1 - t_0) + z_0$, i.e., at point P_1.

Suppose that we knew only the two points P_0 and P_1; with this information, we could still recover the trajectory.

This task is carried out with the *calculus of variations*, here with the principle of stationary action. The idea is that one considers all possible smooth trajectories $z(t)$ (whether real or not) that connect $P_0 = (z_0, t_0)$ and $P_1 = (z_1, t_1)$ and determines the sought-after one through a stationary principle.

To do so, one needs the **Lagrange function** $\mathcal{L}$, which associates a given trajectory $z(t)$ with a function $\mathcal{L}(t)$ of time that depends on the trajectory $z(t)$ and its first time derivative $\dot{z}(t)$:

$$\mathcal{L}(t) = \mathcal{L}(z(t), \dot{z}(t)).$$

Using this function, we can associate a number with each trajectory. This number is the **action**

$$S[z(t)] := \int_{t_0}^{t_1} \mathcal{L}(z(t), \dot{z}(t))\, dt, \tag{15.10}$$

which is nothing more than the time integral of $\mathcal{L}(t)$. The brackets in $S[z(t)]$ indicate that S associates a number with the trajectory $z(t)$ and not with the number $z(t)$. For this reason, S is sometimes called a *functional*.

Now, the *principle of stationary action* says:

> The trajectory $z(t)$ taken by a particle between (z_0, t_0) and (z_1, t_1) is the one for which the action $S[z(t)]$ is stationary.[8]

The solution to this problem can then be given by a differential equation, called the **Euler-Lagrange equation**:

$$\frac{\mathcal{L}}{z} - \frac{d}{dt}\frac{\mathcal{L}}{\dot{z}} = 0.$$

So far, the formalism used is very general. One can use it in optics to find the trajectory of a light ray when it passes through a material with non-constant refractive index (as in the fata morgana) or in mechanics to find the form that a chain assumes when it hangs freely in a uniform gravitational field (catenary). Two further applications are: determining trajectories of mass points that move in force fields (as above) or finding geodesics in curved space.

Particle in the gravitational field. The Lagrange function $\mathcal{L}$ of a particle of mass m in classical mechanics is given by[9] $\mathcal{L} = T - V$, where T is its kinetic energy and V the potential energy. In one dimension, the kinetic energy associated with the particle moving on the trajectory $z(t)$ is given by $T = (m/2)\dot{z}^2$ and, in the gravitational field,

[8] Stationary means that a small change of $z(t)$ does not change the action $S[z(t)]$. This includes the cases when it is extremal, either maximal or minimal.

[9] The force field must be *conservative*, which is the case for the gravitational field in classical mechanics.

the potential energy is given by $V = m\Phi$, where $\Phi(z)$ is the gravitational potential. Therefore,

$$\mathcal{L}(t) = \frac{m}{2}\dot{z}(t)^2 - m\Phi(z(t)), \quad \Phi(z) = gz. \tag{15.11}$$

From that, we have

$$\begin{aligned} \frac{\mathcal{L}}{z} &= -m\Phi'(z) \quad \text{and} \\ \frac{d}{dt}\frac{\mathcal{L}}{\dot{z}} &= \frac{d}{dt}(m\dot{z}) = m\ddot{z} \end{aligned}$$

and the Euler-Lagrange equation gives us

$$m\ddot{z}(t) = -m\Phi'(z) \quad \text{or} \quad \ddot{z}(t) = -\Phi'(z) = -g,$$

which is nothing other than *Newton's force law* (equation of motion).

Therefore: the trajectory of a mass point in the gravitational potential can be described by making the action integral (15.10) over the Lagrange function $\mathcal{L}(t) = T - V$ stationary.

From the principle of stationary action to geodesics. *Can we construct a curved spacetime* (z, t) *whose geodesics are equal to the trajectories of particles that fall freely in the gravitational field?*

Remember that the distance between two close points (z_0, t_0) and (z_1, t_1) in non-curved spacetime (special relativity) is given by

$$\Delta s^2 = c^2\Delta t^2 - \Delta z^2,$$

which is also written in the form

$$\Delta s^2 = g_{00}c^2\Delta t^2 + g_{11}\Delta z^2$$

with the metrics $g_{00} = 1$, $g_{11} = -1$.

In our case, spacetime is almost flat, and we make the Ansatz $g_{00} = 1 + \Delta$, $g_{11} = -1$ for our curved spacetime. The geodesic $z_g(t)$ between $P_0 = (z_0, t_0)$ and $P_1 = (z_1, t_1)$ is the shortest curve $z(t)$ between these two points. Its length is

$$\begin{aligned} s &= \int_{P_0}^{P_1} ds \\ &= \int_{P_0}^{P_1} \sqrt{g_{00}c^2dt^2 - dz^2} \end{aligned}$$

$$
\begin{aligned}
&= \int_{t_0}^{t_1} \sqrt{g_{00}c^2 - \dot{z}^2(t)}\, dt \\
&= \int_{t_0}^{t_1} \sqrt{(1+\Delta)c^2 - \dot{z}^2}\, dt \\
&= \int_{t_0}^{t_1} c\sqrt{1+\Delta - \dot{z}^2/c^2}\, dt.
\end{aligned}
$$

Now, $\dot{z}^2/c^2 \ll 1$, and we will also see that $\Delta \ll 1$. Therefore, we have

$$
s = \int c\sqrt{1+\Delta-\dot{z}^2/c^2}\, dt \approx c\int \left(1 + \frac{\Delta - \dot{z}^2/c^2}{2}\right) dt,
$$

which must become stationary. Now, the additive constant does not matter; it leads to a constant. We can instead make $s - c(t_1 - t_0)$ stationary. Furthermore, a multiplicative constant also does not matter. We multiply $s - c(t_1 - t_0)$ by $-mc$ and get

$$
\bar{s} = \int_{t_0}^{t_1} \left(\frac{m}{2}\dot{z}^2 - \frac{mc^2}{2}\Delta(z)\right) dt.
$$

If we compare this to (15.10) with (15.11), we see that the identification

$$
\Delta(z) = \frac{2\Phi(z)}{c^2}
$$

implies that the trajectory of our mass point in classical mechanics is equal to the geodesic resulting from the calculation at hand.

Therefore: the free motion of a mass point in a curved spacetime with metrics

$$
g_{00}(z) = 1 + \frac{2\Phi(z)}{c^2}, \quad g_{11}(z) = -1
$$

is equivalent to the motion of the mass point in non-curved spacetime under the effect of a uniform gravitational field. In the description with curved spacetime, the gravitation as a force has vanished.

Note that $g_{00}(z)$ is very close to 1, so it can only lead to a very small curvature and trajectories that only slightly deviate from (straight) lines. If we look at the z-ct-diagram (Fig. 15.11, right), we see that the trajectory is indeed very close to a line.

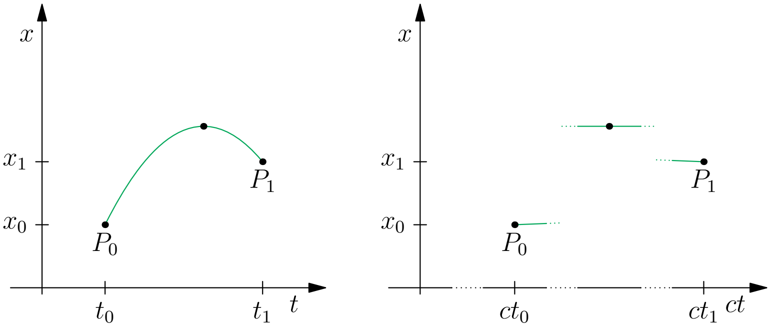


Fig. 15.11 Trajectory of an object in a uniform gravitational field pointing in the negative z-direction. Left: the usual diagram, as drawn in a classical mechanics course. Right: the equivalent diagram with curved spacetime in general relativity. The t-axis is expanded by the factor c

15.5.3 The Complete Picture

In general relativity, all free-falling objects follow geodesics. The effect of gravitation is woven into the geometry of curved spacetime and gravitation as a force disappears.

The metrics is a kind of a gravitational potential. We have shown, how objects move in a given curved spacetime. But how do we come to the curved spacetime (the metric tensor) in the general case?

This task is achieved by **Einstein's field equation**. This is actually a set of 10 coupled non-linear differential equations[10] (i. e., extraordinarily complicated), which, in tensor form, read as

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}. \tag{15.12}$$

To get an idea about how this equation works, look again at Newton's field equation (15.4)

$$\Delta\Phi(\boldsymbol{r}) = 4\pi G \rho(\boldsymbol{r}).$$

Here, on the right side, is the source of the gravitational field, the mass distribution (or density) $\rho(\boldsymbol{r})$. On the left side is a kind of second derivative of the gravitational potential. Given $\rho(\boldsymbol{r})$, with (15.4), we can determine $\Phi(\boldsymbol{r})$ and then, via (15.2), get the gravitational field $\boldsymbol{g}$. The trajectory of a particle in this field is then described by Newton's force law, giving us $\ddot{\boldsymbol{r}} = \boldsymbol{g}$.

In Einstein's field equation (15.12), we have the **energy-momentum tensor** $T_{\mu\nu}$ on the right side. It describes the energy *and* momentum of mass (or energy) distributions. Not only does the mass distribution $\rho(\boldsymbol{r})$ have an influence on spacetime but so does its dynamics. For instance, a rotating sphere generates a different spacetime than a sphere at rest. Something similar occurs in electrodynamics. There, the charge

[10] The indices μ and ν range from 0 to 3. Due to the fact that the tensors are symmetric (i. e., $T_{\mu\nu} = T_{\nu\mu}$), they have only 10 instead of $4 \cdot 4 = 16$ independent components.

density and the charge's velocity (current) also influence the electromagnetic field. The energy-momentum tensor $T_{\mu\nu}$ is a generalization of ρ.

On the right side of Einstein's field equation is the **Einstein tensor** $G_{\mu\nu} := R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}$, which is a function of the metric tensor $g_{\mu\nu}$ and its first and second derivatives.

Then, the situation in Newton's and Einstein's gravity becomes similar: In Newton's gravity, the mass density ρ determines the gravitational potential Φ, and the gravitational field $\boldsymbol{g}$ follows as a derivative from it. Newton's force law $\ddot{\boldsymbol{r}} = \boldsymbol{g}$ then determines the trajectory of an object. In Einstein's gravity, the energy-momentum tensor $T_{\mu\nu}$ determines, via the field equation, the metric tensor $g_{\mu\nu}$, and the Christoffel symbols $\Gamma^{\lambda}_{\mu\nu}$ follow as derivatives from it (the superscript λ is an index, not an exponent). The geodesic equation (the generalization of (15.8) to four-dimensional spacetime) then determines the trajectory of an object.

But there are also large differences. An obvious one is that the formalism of Einstein's gravity is much more difficult than its Newtonian counterpart (this complexity, however, is required and reflects the multitude of facets of gravitation).

Another more conceptual one is that, in Newton's case, space and time exist per se. We can construct a coordinate system, and describe everything with it. In general relativity, however, spacetime is curved, a result of the presence of matter and energy. When we describe the energy-momentum tensor, we do not know how rods and clocks work in curved spacetime, because only the metric tensor $g_{\mu\nu}$ tells us that. For this reason, we have to start with "unphysical" coordinates, which do not correspond to real distances and time differences, but just serve to identify events in spacetime. Solving Einstein's field equation then gives us the metric tensor, which allows us to measure distances and time differences. In most non-trivial cases, this implies an interactive process for solving Einstein's field equations.

15.6 Example: Curved Spacetime Caused by a Large Spherically Symmetric Source

15.6.1 Schwarzschild Metrics

We give a final example, which was the first exact solution to Einstein's field equations, worked out in 1915 by Karl Schwarzschild. It is the gravitational field generated by a mass point of mass M, i.e., about spherically symmetric space. In Newton's gravity, this mass point (imagine the Sun) creates a field with the spherically symmetric gravitational potential given by $\Phi(r) = -GM/r$, which is independent of time.

In general relativity, to determine the metric tensor (which, as we said, is the potential's counterpart), we first incorporate all symmetries of the problem. Using spherical coordinates, $(t, r, \varphi, \vartheta)$, the first fundamental form with the needed symmetries is

$$ds^2 = g_{00}(r,t)c^2\,dt^2 + g_{rr}(r,t)\,dr^2 - r^2(d\vartheta^2 + \sin^2\vartheta\,d\varphi^2).$$

Note that the metric tensor can only depend on the radial coordinate r and the time t. To calculate the tensor components $g_{00}(r,t)$ and $g_{rr}(r,t)$, Einstein's field equations have to be solved. To be able to do this, we would need a great deal of more mathematics than we are using here. We quote the result, which is the **Schwarzschild metrics**:

$$\begin{aligned} g_{00}(r) &= \left(1 - \frac{r_S}{r}\right), \\ g_{rr}(r) &= -g_{00}^{-1}(r,t) = -\left(1 - \frac{r_S}{r}\right)^{-1}. \end{aligned} \tag{15.13}$$

Here, $r_S = 2GM/c^2$ is the **Schwarzschild radius**, which, for the Earth, would be 9 mm and, for the Sun, 3 km, i. e., it is very small. Note that the metric tensor does not depend on the time coordinate.

If the mass point is actually not a point, but rather a spherically symmetric object with a radius r_M, the curved spacetime that it generates outside of the object is equal to that for a mass point. Inside the object, it will be different and, for a fixed $r < r_M$, given by the mass inside of a sphere with radius r.

It is important to remember that the $(t, r, \varphi, \vartheta)$ in Schwarzschild's solution are only Gaussian coordinates and Δr and/or Δt are not real distance or time intervals. Only if $M = 0$, i. e., $r_S = 0$, do the coordinates refer to real distances and time intervals. The coordinates also become spherical coordinates for very large r, because, for $r \to \infty$, $g_{00} \to 1$ and $g_{rr} \to -1$.

The *local time* at a fixed coordinate r_0 is given by $dr = d\varphi = d\vartheta = 0$, i.e.,

$$d\tau = \sqrt{g_{00}(r)}\,dt = \sqrt{1 - \frac{r_S}{r}}\,dt = \sqrt{1 + \frac{2}{c^2}\Phi(r)}\,dt,$$

where $\Phi(r)$ is Newton's gravitational potential. This is exactly in line with what we got from the Gedanken experiment of the free-falling elevator.

In comparison with the clock at $r = \infty$ (referring to "very far away from the central mass"), which proceeds according to coordinate time t, the local time proceeds more slowly. At $r = r_S$, the metrics has a singularity. The closer the clock at $r > r_S$ comes to $r = r_S$, the more slowly it proceeds (in comparison to the clock at $r = \infty$), and, at $r = r_S$, it comes to a rest. This means that objects (massive objects, light) at $r \leq r_S$ can never reach locations with $r > r_S$. The hypersurface $r = r_S$ is called the **event horizon**, and if the central object with mass M has a radius smaller than r_S, it is a **black hole**.

A person that freely falls in a cabin will not notice when the cabin crosses the event horizon. All experiments in the cabin will work as expected and a measurement of the speed of light will yield c.

Local distances at a fixed time and in the radial direction have $dt = d\varphi = d\vartheta = 0$, and therefore

$$dl = \sqrt{-ds^2} = \sqrt{-g_{rr}(r)}\,dr = \sqrt{1 - \frac{r_S}{r}}^{\,-1} dr = \sqrt{\frac{r}{r - r_S}}\,dr.$$

In comparison to distances at $r = \infty$, where they are given by coordinate differences, local distances are larger. The closer to the event horizon, the larger the distances. Note again the singularity in g_{rr} at $r = r_S$.

For an observer at $r = \infty$, a light ray at (r, t) has the velocity

$$c(r) = \frac{dr}{dt} = \frac{\sqrt{g_{00}}}{\sqrt{-g_{rr}}} = g_{00} \cdot c = \sqrt{1 - \frac{r}{r_S}} \cdot c.$$

The closer the light ray is to $r = r_S$, the more slowly it proceeds.

15.6.2 The Embedding Diagram

There is a nice procedure for visualizing two-dimensional curved spaces in three-dimensional flat space, which we can apply to the Schwarzschild metrics (15.13). First, we have to reduce the Schwarzschild metrics to two dimensions. Due to the fact that the metrics is independent of the time, we can just set $t = 0$, which removes the time dimension. Furthermore, the metrics has rotational symmetry and, without any loss, we can choose the subspace (r, φ) by setting $\vartheta = \pi/2$. This leaves us with the two-dimensional curved space with the metrics

$$-\,ds^2 = -g_{rr}(r)\,dr^2 + r^2\,d\varphi^2. \tag{15.14}$$

Now, we add a third dimension with coordinate z that is perpendicular to the (r, φ)-plane and get cylinder coordinates as Gaussian coordinates. We construct a surface with the property that the distance from (r, φ) to $(r + dr, \varphi + d\varphi)$ in the curved space (15.14) is equal to the distance between the two points (r, φ, z) and $(r + dr, \varphi + d\varphi, z + dz)$, where $z = z(r, \varphi)$ and $z + dz = z(r + dr, \varphi + d\varphi)$, on the surface in the three-dimensional embedding space (Fig. 15.12).

This means that we have to construct a surface $z(r, \varphi)$ in our three-dimensional flat space with cylinder coordinates such that

$$\begin{aligned} -ds^2 &= dz^2 + dr^2 + r^2\,d\varphi^2 \\ &= \left[\left(\frac{dz}{dr}\right)^2 + 1\right] dr^2 + r^2\,d\varphi^2. \end{aligned}$$

This requires

$$\left(\frac{dz}{dr}\right)^2 + 1 = -g_{rr}(r) = \left(1 - \frac{r_S}{r}\right)^{-1},$$

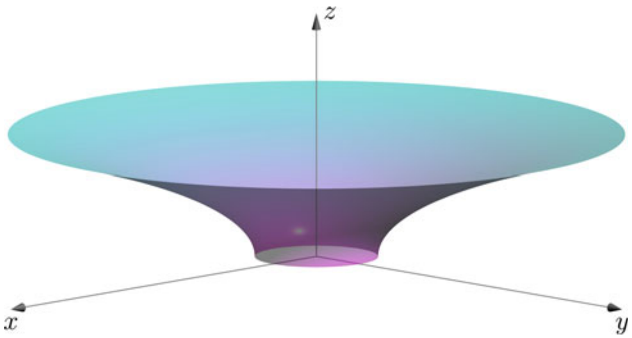


Fig. 15.12 Embedding a two-dimensional curved subspace of spacetime with Schwarzschild metrics into three-dimensional flat space

from which we get

$$\frac{dz}{dr} = \pm\sqrt{\frac{r_S}{r - r_S}} \quad \text{and} \quad z(r) = \pm 2\sqrt{r_S(r - r_S)}.$$

We can turn this around and have $r(z) = z^2/(4r_S) + r_S$, which is a parabola (called Flamm's paraboloid) rotated around the z-axis. The distance when we want to go from point $P_1 = (r_1, \varphi_1)$ to point $P_2 = (r_2, \varphi_2)$ in curved two-dimensional Schwarzschild space is equal to the distance from $Q_1 = (r_1, \varphi_1, z_1)$ with $z_1 = z(r_1)$ to $Q_2 = (r_2, \varphi_2, z_2)$ with $z_2 = z(r_2)$ on the curved surface in flat three-dimensional space.

Chapter 16
Summary

The **central pillars of special relativity** are its two principles:

- **Einstein's principle of relativity**: all inertial frames are on an equal footing.
- **Principle of the absolute speed of light**: light propagates with the same universal speed c in all directions and all inertial frames.

Einstein synchronization also plays an important role, and, arguably, one could adopt a different synchronization scheme. The main effects predicted by the theory, however, would be the same with a different synchronization scheme.[1]

Out of the two principles, the relativity principle is that with the more drastic consequences. It implies that either classical mechanics or electrodynamics cannot be exactly correct.

No more than the two principles are needed to derive the "*kinetic effects*": *relativity of simultaneity*, *length contraction* and *time dilation*. These are real effects, probed innumerable times in experiments. Along with these effects, we must use the *Lorentzian addition of velocities*, the usual addition of velocities being insufficient.

The transformation that mediates between two inertial frames is the *Lorentz transformation*. All kinematic effects that are also a consequence of the two principles follow from the Lorentz transformation.

The laws of classical mechanics cannot be used anymore because they are not form-invariant under Lorentz transformation and, if correct, would predict differences between inertial frames and contradict Einstein's principle of relativity. There must be a *new mechanics* that is (a) form-invariant under Lorentz transformation and (b) become classical mechanics within the limit of small velocities. This is Einstein's mechanics, a part of special relativity.

These requirements lead to the definition of the *(relativistic) energy* and the *(relativistic) momentum*. The (relativistic) energy involves the famous formula $E = mc^2$:

[1] An example is Lorentz's ether theory, which could be seen as a theory that is very similar to Einstein's theory, with the exception that it uses an absolute time.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5_16

the rest energy and the mass of an object are the same. Conservation laws for the (relativistic) energy and momentum are form-invariant under LT, which is again required by Einstein's principle of relativity.

Maxwell's *electrodynamics as is* is form-invariant under Lorentz transformations; there's no need to adapt it to fit with the principles of special relativity. Contrary to what the fathers of electrodynamics thought, it is valid in the same form in each inertial frame and there is no luminiferous aether. The electromagnetic fields transfer into each other when changing from one to another inertial frame.

Special relativity has ample relevance for our daily life: *satellite navigation would not work* without Einstein's theories of relativity.

Einstein's achievement was made "on the shoulders" of colleagues. But, in the end, it was his genius that understood the profound consequences and that carried out the inevitable revolution.

As Darrigol [Darrigol05] put it:

> "Most of the components of Einstein's paper appeared in others' anterior works on the electrodynamics of moving bodies. Poincaré and Alfred Bucherer had the relativity principle. Lorentz and Larmor had most of the Lorentz transformations, Poincaré had them all. Cohn and Bucherer rejected the ether. Poincaré, Cohn, and Abraham had a physical interpretation of Lorentz's local time. Larmor and Cohn alluded to the dilation of time. Lorentz and Poincaré had the relativistic dynamics of the electron. None of these authors, however, dared to reform the concepts of space and time. None of them imagined a new kinematics based on two postulates. None of them derived the Lorentz transformations on this basis. None of them fully understood the physical implications of these transformations. It all was Einstein's unique feat."

Appendix
Useful Formulas

A.1 Frequently Used Approximations

For $|x| \ll 1$, the following approximations are useful:

$$\begin{aligned} \frac{1}{1 \pm x} &\approx 1 \mp x, \\ \sqrt{1-x} &\approx 1 - \frac{1}{2}x, \\ \frac{1}{\sqrt{1-x}} &\approx 1 + \frac{1}{2}x, \\ \sin x \approx x & \qquad (x \text{ in radians}), \\ \tan x \approx x & \qquad (x \text{ in radians}). \end{aligned}$$

For $\gamma(u)$ from (8.3), for the case $|u| \ll c$, the following holds:

$$\gamma^{-1}(u) \approx 1 - \frac{u^2}{2c^2}, \quad \gamma(u) \approx 1 + \frac{u^2}{2c^2}.$$

A.2 From Special Relativity

A.2.1 The Doppler Effect

$$\frac{\nu_{\mathrm{O}}}{\nu_{\mathrm{S}}} = \frac{1}{\gamma_v \cdot (1 + \boldsymbol{v}\boldsymbol{e}/c)}.$$

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5

Here: $\boldsymbol{v}$ is the velocity of the light source and $\boldsymbol{e}$ its direction, both at the moment when the light pulse was emitted and in the observer's frame. See (9.6).

A.2.2 Aberration

Suppose Bob moves relative to Alice with velocity v in the direction of their common x-axis. Alice sees the light of a star coming from the direction φ, measured from the x-axis. Then, Bob sees the light from this star coming from the direction φ', where

$$\tan\varphi' = \frac{1}{\gamma_v}\frac{\sin\varphi}{\cos\varphi + v/c}$$

(see 12.9). Or, alternatively, as in (12.10),

$$\cos\varphi' = \frac{\cos\varphi + v/c}{1 + (v/c)\cos\varphi}.$$

A.2.3 Lorentzian Addition of Velocities

From (12.6), it follows that

$$u'_{\parallel} = \frac{1}{1 - \frac{\boldsymbol{u}\boldsymbol{v}}{c^2}}(u_{\parallel} - v)$$

$$\boldsymbol{u}'_{\perp} = \gamma_v^{-1}\frac{1}{1 - \frac{\boldsymbol{u}\boldsymbol{v}}{c^2}}\boldsymbol{u}_{\perp}.$$

If $\boldsymbol{v} = v\boldsymbol{e}_v$ and we denote with $\hat{P}_{\boldsymbol{e}_v}$ the projector on $\boldsymbol{e}_v$, we can write:

$$\boldsymbol{u}' = \boldsymbol{u} \ominus \boldsymbol{v} = \frac{(\hat{P}_{\boldsymbol{e}_v}\boldsymbol{u} - \boldsymbol{v}) + \gamma_v^{-1}(\boldsymbol{u} - \hat{P}_{\boldsymbol{e}_v}\boldsymbol{u})}{1 - \boldsymbol{u}\boldsymbol{v}/c^2} = \frac{(\hat{P}^{\parallel}_{\boldsymbol{e}_v}\boldsymbol{u} - \boldsymbol{v}) + \gamma_v^{-1}(\hat{P}^{\perp}_{\boldsymbol{e}_v}\boldsymbol{u})}{1 - \boldsymbol{u}\boldsymbol{v}/c^2}.$$

A.2.4 Others

Decomposition of a vector $\boldsymbol{r}$ into a component $\boldsymbol{r}_{\parallel}$ parallel to a unit vector $\boldsymbol{e}$ and another component $\boldsymbol{r}_{\perp}$ perpendicular to $\boldsymbol{e}$:

$$\boldsymbol{r} = \boldsymbol{r}_\parallel + \boldsymbol{r}_\perp :$$
$$\boldsymbol{r}_\parallel = (\boldsymbol{r}\boldsymbol{e})\boldsymbol{e}$$
$$\boldsymbol{r}_\perp = \boldsymbol{r} - (\boldsymbol{r}\boldsymbol{e})\boldsymbol{e} = -(\boldsymbol{r} \times \boldsymbol{e}) \times \boldsymbol{e}.$$

γ-formulas (see 12.5):

$$\gamma_{u'} u' = \gamma_v \gamma_u (u - v)$$
$$\gamma_{u'} = \gamma_v \gamma_u (1 - uv/c^2).$$

List of Sources

- Fig. 2.2: Figure reproduced from W. Bertozzi: *Speed and Kinetic Energy of Relativistic Electrons*, Am. J. Phys. **32**(7), 551 (1964) with the permission of the American Association of Physics Teachers.
- Fig. 4.1: Right: Wikipedia, by Thomas Reisinger, https://commons.wikimedia.org/wiki/File:Poissonspot_simulation_d4mm.jpg, license: CC BY-SA 3.0.
- Fig. 4.17: Wikipedia, by Takasunrise0921, https://commons.wikimedia.org/wiki/File:Prime-meridian.jpg, license: CC BY-SA 3.0.
- Fig. 4.19: Own work. Idea taken from Scheffler/Elsässer, Bau und Physik der Galaxis, p. 57.
- Fig. 5.4: Wikipedia, by boson, http://commons.wikimedia.org/wiki/File:Michelsonnachbau.jpg, license: CC-BY-SA-2.5.
- Fig. 5.5: Left: A A Michelson, E Morley: *On the Relative Motion of the Earth and the Luminiferous Ether*, Am. J. Science **34**, 333 (1887). See https://commons.wikimedia.org/wiki/File:On_the_Relative_Motion_of_the_Earth_and_the_Luminiferous_Ether_-_Fig_3.png, license: public domain.
- Fig. 5.5: Right: dto. See https://commons.wikimedia.org/wiki/File:On_the_Relative_Motion_of_the_Earth_and_the_Luminiferous_Ether_-_Fig_4.png, license: public domain
- Fig. 5.6: A Michelson, E Morley: *On the Relative Motion of the Earth and the Luminiferous Ether*, Am. J. Science **34**, 333 (1887). See https://commons.wikimedia.org/wiki/File:On_the_Relative_Motion_of_the_Earth_and_the_Luminiferous_Ether_-_Fig_6.png, license: public domain.
- Fig. 7.2: Wikipedia, by Andreas Schwarzkopf, https://commons.wikimedia.org/wiki/File:Klaus_Rinke_Zeitfeld,_1987.jpg, license: CC BY-SA 4.0.
- Fig. 7.11: Roulette: source: https://www.freepik.com/free-vector/vector-realistic-casino-roulette-wheel-side-view-isolated-green-poker-table_11062553.htm, attribution: “Designed by macrovector/Freepik”, license: free.

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5

- Fig. 8.7: Own work. Idea taken from Wikipedia, figures by Youyz, https://commons.wikimedia.org/wiki/File:Ladder_Paradox_GarageScenario.svg and https://commons.wikimedia.org/wiki/File:Ladder_Paradox_LadderScenario.svg, both in public domain.
- Fig. 8.11: Left: Figure reproduced from Gamov's book "Mr Tompkins in Wonderland" with the permission of The Licensor through PLSclear. © New York: The Macmillan Company; Cambridge, Eng.: The University Press, 1940.
- Fig. 8.15: Figure reproduced from https://www.tempolimit-lichtgeschwindigkeit.de/bewegung_d/tempolimit-material.pdf with the permission of U. Kraus.
- Fig. 8.16: Figure reproduced from U. Kraus *et al.*: *Was Einstein noch nicht sehen konnte*, Physik-Journal **7**, 1 (2002) with the permission. © Wiley-VCH GmbH.
- Fig. 9.5: Wikipedia, by Luekk, https://commons.wikimedia.org/wiki/File:GB-3-Gew-Pendeluhr_(Luekk).jpg, license: CC-BY-SA 3.0.
- Fig. 9.5: Wikipedia, by Chris Burks (Chetvorno), https://commons.wikimedia.org/wiki/File:Pocket_Watch_Balance_Wheel_a.JPG, license: public domain.
- Fig. 9.5: Wikipedia, by Chribbe76, https://commons.wikimedia.org/wiki/File:Inside_QuartzCrystal-Tuningfork.jpg, license: public domain.
- Fig. 9.6: Wikipedia, by Jörg Behrens, http://commons.wikimedia.org/wiki/File:Atomuhr-CS2.jpg, license: CC BY-SA 3.0. Cut to size.
- Fig. 9.21: Figure reproduced from https://www.leifiphysik.de/relativitaetstheorie/spezielle-relativitaetstheorie/ausblick/zwillingsparadoxon, see http://www.leifiphysik.de/sites/default/files/medien/zwilinge_spezrelatheorie_aus.jpg, license: "Fair use" applies because copyright holder was not identifiable.
- Fig. 9.25: Figure reproduced from http://www.lightandmatter.com/article/hafele_keating.html, license: "Fair use" applies because copyright holder was not identifiable.
- Fig. 9.28: Wikipedia, by Trex2001, https://commons.wikimedia.org/wiki/File:GPS_Spheres.svg, license: CC BY-SA 3.0. Figure was minimally modified.
- Fig. 15.6: Wikipedia, by Eric Gaba (Sting), https://commons.wikimedia.org/wiki/File:Minimal_surface_curvature_planes-en.svg, license: CC BY-SA 3.0.
- Fig. 15.7: Wikipedia, by Cmglee, https://commons.wikimedia.org/wiki/File:Comparison_of_geometries.svg. license: CC BY-SA 4.0. Labeling of figure was removed.

Solution to Exercises

Solution to Exercise 15: We start with the direction perpendicular to the relative motion. Let $T_\perp$ be the time light needs to travel from the beam splitter to the mirror. During this time, the interferometer has moved by $vT_\perp$, and for the length $l_\perp$ of the trajectory, we have $l_\perp^2 = L^2 + (vT_\perp)^2 = L^2 + (v/c)^2 l_\perp^2$. Therefore,

$$T_\perp = \frac{2l_\perp}{c} = \frac{2L}{c}\frac{1}{\sqrt{1-\beta^2}},$$

which is the same as (5.3).

Now, we come to the direction parallel to the relative motion. Let $T_{\|+}$ be the time the light needs for the trajectory from the beam splitter to the mirror. During this time, the mirror moves away by the distance $vT_{\|+}$, so the actual traveling distance is $l_{\|+} = L + vT_{\|+}$ and the time needed is $T_{\|+} = l_{\|+}/c$ or

$$T_{\|+} = \frac{L}{c}\frac{1}{1-\beta}.$$

When the light comes back, the beam splitter moves toward it, and with the same arguments, we get

$$T_{\|-} = \frac{L}{c}\frac{1}{1+\beta}.$$

This leads to (5.2), and eventually to (5.4).

Solution to Exercise 17: Suppose T_{Ecl} is the time between two eclipses of Io. In this time, Io performs a complete orbit around Jupiter, plus a small angle α. In the same time, Jupiter moves forward by this small angle α on its orbit around the Sun. We have

$$T_{\text{Ecl}} = \left(1 + \frac{\alpha}{2\pi}\right) T_{\text{Io}} = \frac{\alpha}{2\pi} T_{\text{Jup}}.$$

From this, we get

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5

$$\begin{aligned}\alpha &= 2\pi \cdot \frac{T_{\text{Io}}}{T_{\text{Jup}} - T_{\text{Io}}}\\ &= 2\pi \cdot \frac{1.769}{365.2425 \cdot 11.86 - 1.769}\\ &= 2.567 \times 10^{-3}\end{aligned}$$

and

$$\Delta T = T_{\text{Ecl}} - T_{\text{Io}} = \frac{\alpha}{2\pi} T_{\text{Io}} = 63.4\,\text{s}.$$

The mean distance between the Earth and the Sun is about 150 million kilometers, therefore, light needs $2 \times 1.5 \times 10^8/(3 \times 10^5)\,\text{s} = 1000\,\text{s}$.

In Exercise 23 at the end of Sect. 7.9.2, you find a very similar problem.

Solution to Exercise 19: Let a light pulse start at clock A at $t_{A,0}$ (see Fig. 7.4). It will arrive at the semitransparent mirror M, where it is split. The reflected pulse then arrives at clock A at $t_{A,1}$ and the transmitted one at clock B at $t_{B,1}$. The latter clock sends the pulse immediately back to clock A, where it arrives at $t_{A,2}$ (we neglect the semitransparent mirror now).

Suppose that *Einstein synchronization* has been performed. Then, by definition, $t_{B,1} = (t_{A,0} + t_{A,2})/2$ or $t_{B,1} = t_{A,0} + \Delta t$ with $\Delta t = (t_{A,2} - t_{A,0})/2$. Therefore, the light pulse needs the time Δt to travel from clock B back to clock A. Due to the fact that the semitransparent mirror is halfway between the clocks, the light pulse from the semitransparent mirror to clock A will take $\Delta t/2$, exactly the same as the light pulse from the semitransparent mirror to clock B. Therefore, if two clocks are Einstein synchronized, they are also synchronized according to symmetric synchronization.

Suppose now that the *symmetric synchronization* procedure has been carried out in the following slightly modified way. The light pulse starts at A at $t_{A,0}$ and is split at M at t_M, where the actual synchronization starts. The times at which the split pulses arrive at A and B are $t_{A,1}$ and $t_{B,1}$, respectively. Let us set $\Delta t = t_{A,1} - t_{A,0}$. Now, due to the fact that the path from clock A to clock B is twice as long as the path from clock A to the semitransparent mirror, $t_{A,2}$ is given by $t_{A,2} = t_{A,0} + 2\Delta t$, and we get $(t_{A,2} + t_{A,0})/2 = t_{A,0} + \Delta t = t_{A,1}$. On the other hand, this must be $t_{B,1}$ according to Einstein synchronization. Therefore, if two clocks are synchronized according to symmetric synchronization, they are also Einstein synchronized.

This completes the proof that symmetric synchonization is equivalent to Einstein synchronization.

Solution to Exercise 20: The coordinates of the “flash” events are:

$$E_1\colon \left(t_1, x_1 = \frac{c^2}{v} t_1\right), \quad E_2\colon \left(t_2, x_2 = \frac{c^2}{v} t_2\right).$$

Suppose that $t_2 > t_1$.

For the world lines of the light pulses, we get

$$\begin{aligned} L_1\colon \quad (x - x_1) &= +c(t - t_1) \\ x &= +ct - ct_1 + x_1 = +ct + \left(\frac{c^2}{v} - c\right) t_1, \\ L_2\colon \quad (x - x_2) &= -c(t - t_2) \\ x &= -ct + ct_2 + x_2 = -ct + \left(\frac{c^2}{v} + c\right) t_2. \end{aligned}$$

The intersection event E_{I} is given by $E_{\mathrm{I}} = L_1 \cap L_2$. Therefore,

$$\begin{aligned} ct_{\mathrm{I}} + \left(\frac{c^2}{v} - c\right) t_1 &= -ct_{\mathrm{I}} + \left(\frac{c^2}{v} + c\right) t_2, \\ 2ct_{\mathrm{I}} &= c(t_1 + t_2) + \frac{c^2}{v}(t_2 - t_1), \\ t_{\mathrm{I}} &= \frac{t_1 + t_2}{2} + \frac{c}{v}\frac{t_2 - t_1}{2}. \end{aligned}$$

E_{I} lies on L_1, and therefore

$$\begin{aligned} x_{\mathrm{I}} &= c(t_{\mathrm{I}} - t_1) + x_1 \\ &= c\frac{t_1 + t_2}{2} + \frac{c^2}{v}\frac{t_2 - t_1}{2} - ct_1 + \frac{c^2}{v} t_1 \\ &= c\frac{t_2 - t_1}{2} + \frac{c^2}{v}\frac{t_1 + t_2}{2}. \end{aligned}$$

The line through E_{I} and parallel to the x'-axis is given by

$$\begin{aligned} x - x_{\mathrm{I}} &= v(t - t_{\mathrm{I}}) \\ x &= vt - vt_{\mathrm{I}} + x_{\mathrm{I}} \\ &= vt + \left(\frac{c^2}{v} - v\right) \frac{t_1 + t_2}{2}. \end{aligned}$$

The intersection $E_{\mathrm{M}} = (t_{\mathrm{M}}, x_{\mathrm{M}})$ with Bob’s x'-axis, given by $x = (c^2/v)t$, is

$$\frac{c^2}{v}t_{\mathrm{M}} = vt_{\mathrm{M}} + \left(\frac{c^2}{v} - v\right)\frac{t_1 + t_2}{2} \qquad \Longrightarrow \qquad t_{\mathrm{M}} = \frac{t_1 + t_2}{2}.$$

Furthermore,

$$x_{\mathrm{M}} = \frac{c^2}{v}t_{\mathrm{M}} = \frac{c^2}{v}\frac{t_1 + t_2}{2} = \frac{x_2 + x_1}{2}.$$

The event E_{M} in the spacetime diagram is in the middle between E_1 and E_2, and therefore, for Bob, the light pulses from E_1 and E_2 that were emitted at $t' = 0$ at x'_1 and x'_2, respectively, meet at $x'_{\mathrm{M}} = (x'_1 + x'_2)/2$.

Solution to Exercise 23: We give two methods.

In the first method, we calculate the position of the clock's hands at time t and require these to be equal. Measured from the vertical, the clockwise measured angle φ_S of the hour hand is given by $\varphi_S(t) = 2\pi(t/12\,\mathrm{h} - n_S)$, where n_S is the number of complete laps. For the minute hand, we have $\varphi_M(t) = 2\pi(t/1\,\mathrm{h} - n_M)$. When the clock hands coincide for the first time, we have $n_S = 0$ and $n_M = 1$. From that, we have $t/12\,\mathrm{h} \overset{!}{=} t/1\,\mathrm{h} - 1$ or $t \cdot (1 - 1/12) = 1\,\mathrm{h}$, and therefore $t = (12/11)\,\mathrm{h} = 1\,\mathrm{h}\,5\,\mathrm{min}\,27.27\,\mathrm{s}$.

The second method is much easier: the clock's hands coincide exactly 11 times in 12 h. Therefore, this happens every $(11/12)\,\mathrm{h}$.

In the discussion of the Sagnac interferometer, we have given the time Δt by l/c. This is only approximately correct, because, while the light traveled the distance from l to $l + \Delta l$, the Earth has kept rotating. So, to be really exact, we have to write $\Delta t = l/(c - \Omega R)$, but due to the fact that the velocity $v_R = \Omega R$ of the loop usually is much smaller than the speed of light, one can neglect this.

Solution to Exercise 29: We start with the velocity v_0 of the object at $t = t_0$. Derivation of the hyperbola yields $x(t)\dot{x}(t) = c^2 t$, and therefore $v_0 = \dot{x}(t_0) = c^2 t_0/x(t_0) = c^2 t_0/x_0$. Therefore, the t'-axis is given by $x = v_0 t$ and the x'-axis by $x = (c^2/v_0)t$. Consequently, the event (t_0, x_0) lies on the x'-axis; in other words, at $t' = 0$, the object's velocity vanishes.

The Lorentz transformation is $x = \gamma(v_0)(x' + v_0 t')$, $t = \gamma(v_0)(t' + (v/c^2)x')$. Plugging this into the formula of the hyperbola yields $x'^2 - c^2 t'^2 = c^4/\alpha^2$. Therefore, the acceleration in all instantaneous rest frames of the object is the same, or, to put it another way, the proper acceleration of the object is constant.

Solution to Exercise 32: We describe the process from the point of view of Alice. Let l_0 be the proper length of the light clock. If said clock is oriented in Bob's direction of motion, for Alice, it is contracted and has the length $l_{\mathrm{A},\parallel} = l_0/\gamma_v$. Thus, one clock period for Alice takes $t_0 = t_\parallel = (2l_{\mathrm{A},\parallel}/c)/(1 - v^2/c^2) = 2\gamma_v^2 l_{\mathrm{A},\parallel}/c$ (compare to 5.2).

If the clock is oriented perpendicular to the direction of motion, for Alice, it has the (so far unknown) length $l_{\mathrm{A},\perp}$. The period again must be t_0. Thus, for Alice, the light pulse in the clock must cover a distance of $2\sqrt{(vt_0/2)^2 + l_{\mathrm{A},\perp}^2}$. Consequently, we have $(ct_0/2)^2 = (vt_0/2)^2 + l_{\mathrm{A},\perp}^2$ or $t_0 = (2l_{\mathrm{A},\perp}/c)/\sqrt{1 - v^2/c^2}$ (compare to 5.3).

Equating both expressions yields $l_{A,\perp} = \gamma_v l_{A,\parallel} = \gamma_v(l_0/\gamma_v) = l_0$. Therefore, dimensions perpendicular to the direction of motion are not contracted!

Solution to Exercise 35: Deriving the hyperbola $c^2t^2 - x^2 = 1$ with respect to the time yields $\dot{x} = c^2t/x$. Inserting $x = vt$ then yields the slope of the t'-axis.

Solution to Exercise 37: For the relativistic Doppler effect, we have

$$\frac{\nu_O}{\nu_S} = \sqrt{\frac{1 - v/c}{1 + v/c}} \approx \sqrt{\left(1 - \frac{v}{c}\right)^2} = 1 - \frac{v}{c},$$

while, for the classical Doppler effect, one gets

$$\frac{\nu_O}{\nu_S} = \frac{1 - v_O/c}{1 - v_S/c} \approx (1 - v_O/c)(1 + v_S/c) \approx (1 - (v_S - v_O)/c) = 1 - v/c.$$

Here, $v = v_S - v_O$ is the velocity of the source relative to the observer.

Solution to Exercise 39: This is easy. Under the integral, only the square of the velocity appears, and this is constant. Therefore, from (9.13), we have

$$\Delta t' = \int_{t_Q}^{t_R} \sqrt{1 - v_B^2(t)/c^2}\, dt = \gamma_v^{-1} \int_{t_Q}^{t_R} dt = \gamma_v^{-1} \Delta t.$$

Solution to Exercise 40: Light covers a distance of 1 m in about 3 ns. The relative error would be approximately 10^{-12}. Therefore, it would take only about 300 s, or five minutes. For this reason, it is necessary that satellite clocks be corrected regularly through the use of commands sent by control stations on the Earth. Taking into account this regular correction, the satellite clocks stay synchronized with the clocks on the Earth.

Solution to Exercise 44: This follows directly from the relation

$$\tanh(\alpha + \beta) = \frac{\tanh\alpha + \tanh\beta}{1 + \tanh\alpha\tanh\beta}.$$

Solution to Exercise 45: Let Alice and Bob be inertial observers. Bob moves with velocity v in the x-direction relative to Alice. There are two clocks at equal distances in front of and behind him, both of which have to be synchronized (for Bob). At time $t = t' = 0$, Bob sends a signal with velocity u toward the clocks. When the signals arrive at the clocks, those are set to zero. Then, Bob assumes these clocks to be synchronized.

How does Alice describe this experiment? For Alice, Bob moves on the trajectory $x = vt$, while the trajectories of the clocks are given by $x = vt \pm L/2$. The signals obviously have different velocities for Alice than for Bob. Let $x = u_+t$ be the trajectory of the light pulse traveling to the front clock and $x = u_- t$ the one traveling to the rear clock. In the case of classical mechanics, $u_\pm = v \pm u$, whereas, in special relativity, (10.3) holds, and we therefore have $u_\pm = (v \pm u)/(1 \pm uv/c^2)$.

At event E_+, the light signal arrives at the front clock, and at event E_-, it arrives at the rear clock. We calculate Alice's coordinates at these events. The time coordinates $t_\pm$ follow from the intersection of $x = u_\pm t$ and $x = vt \pm L/2$. This yields

$$t_\pm = \frac{\pm L}{2(u_\pm - v)}.$$

In the case of classical mechanics, $u_\pm - v = \pm u$, and therefore $t_\pm = L/(2u)$. The events therefore are also simultaneous for Alice. This is what we expected, because, in classical physics, simultaneity is absolute. Events, that are simultaneous for Bob, are also simultaneous for Alice.

In the case of special relativity, the time coordinates of the two events are different. Bob's axis of simultaneity goes through both events. We calculate the slope of the line connecting these events.

First, we have $x_\pm = vt_\pm \pm L/2$. From that,

$$\frac{\Delta x}{\Delta t} = \frac{x_+ - x_-}{t_+ - t_-} = \frac{v(t_+ - t_-) + L}{t_+ - t_-} = v + \frac{L}{t_+ - t_-}$$

follows. The denominator in the expression for the time coordinates is

$$u_\pm - v = \frac{v \pm u}{1 \pm uv/c^2} = \frac{\pm u(1 - v^2)}{1 \pm uv/c^2}.$$

From that,

$$\begin{aligned}\Delta t = t_+ - t_- &= \frac{L}{2}\left(\frac{1}{u_+ - v} - \frac{1}{u_- - v}\right) = \frac{L}{2}\frac{(1 + uv/c^2) - (1 - uv/c^2)}{u(1 - v^2)}\\ &= \frac{L}{2}\frac{2uv}{u(1 - v^2)c^2} = L \cdot \frac{v}{(1 - v^2)c^2}\end{aligned}$$

and

$$\frac{\Delta x}{\Delta t} = v + \frac{L}{t_+ - t_-} = v + c^2\frac{1 - v^2}{v} = \frac{c^2}{v}$$

follows.

Here, the signal velocities u_+, u_- do not appear anymore. The velocity of the signals used to synchronize events does not matter as long as one pays attention to the relativistic addition of velocities. The result is the same: events that, for Alice, lie on a line with slope c^2/v are simultaneous for Bob. For signals with the speed of light, however, the synchronization procedure is much easier.

Solution to Exercise 50: That $(a_0, \boldsymbol{a})$ is a four-vector means that, in the case of a Lorentz transformation (11.9), it transforms as

$$a'_0 = \gamma_v(a_0 - (v/c)a_x), \quad a'_x = \gamma_v(a_x - (v/c)a_0), \quad a'_y = a_y, \quad a'_z = a_z.$$

The same holds for $(b_0, \boldsymbol{b})$. Now it's just a matter of plugging these formulas into $a'_0b'_0 - \boldsymbol{a}'\boldsymbol{b}'$ and showing that this is equal to $a_0b_0 - \boldsymbol{ab}$. We leave the details to the reader.

Solution to Exercise 51:

(b) Suppose the rod in Σ has length ξ_0. Then, its length in S, which is to be determined, is x_0. This is the x-coordinate of event E_0. Its t-coordinate is $t_0 = 0$, and from that follows $\tau_0 = d\xi_0$. Now, we plug (τ_0, ξ) into the transformation (11.11) and get

$$x_0 = b \cdot (1 - dv)\xi_0.$$

In special relativity, this gives us $x_0 = \xi_0/\gamma$, which is length contraction. For the moving observer, the rod at rest in Σ is shorter than for the observer at rest in Σ.

(c) We need to apply simultaneity for S. The t-coordinate of E_0 must be equal to the t-coordinate of $(\tau_0, 0)$, therefore, $t_0 = a \cdot (\tau_0 - d \cdot 0)$ or

$$t_0 = a\tau_0.$$

In special relativity, we have $t_0 = \gamma\tau_0$, which means time dilation. The moving observer sees the clock at rest in Σ run slow.

(d) The front end of the rod is at $x = x_0$. From (11.11), this gives us $\xi = v\tau + x_0/b$, and for $\tau = 0$, finally,

$$\xi_0 = x_0/b.$$

In special relativity, we have $\xi_0 = x_0/\gamma$, which is is length contraction again. The observer at rest in Σ sees the moving rod contracted.

(e) From the transformation, we directly get $t_0 = a \cdot (\tau_0 - d\xi_0) = a \cdot (1 - dv)\tau_0$ or

$$\tau_0 = \frac{1}{a \cdot (1 - dv)} t_0.$$

In special relativity, this gives us $\tau_0 = \gamma t_0$, i. e., for the observer in Σ, their own clock runs faster than the moving clock. This again is time dilation.

This shows that a or $1/(a \cdot (1 - dv))$, respectively, is responsible for time dilation. Furthermore, $1/b$ or $b(1 - dv)$, respectively, is responsible for length contraction.

If the relativity principle holds, time dilation must be the same in both reference frames, and therefore $a^2 = 1/(1 - dv)$. The same holds for length contraction, and therefore $b^2 = 1/(1 - dv) = a^2$.

Solution to Exercise 52: We use matrix notation and set $c = 1$. Then,

$$\begin{pmatrix} t' \\ x' \end{pmatrix} = \hat{L}(v) \begin{pmatrix} t \\ x \end{pmatrix}, \begin{pmatrix} t'' \\ x'' \end{pmatrix} = \hat{L}(v') \begin{pmatrix} t' \\ x' \end{pmatrix} \quad \text{with} \quad \hat{L}(v) = \gamma_v \begin{pmatrix} 1 & -v \\ -v & 1 \end{pmatrix}.$$

If we transform first from Alice's coordinates to Bob's coordinates and then from Bob's coordinates to Claire's coordinates, we get

$$\begin{pmatrix} t'' \\ x'' \end{pmatrix} = \hat{L}(v')\hat{L}(v) \begin{pmatrix} t \\ x \end{pmatrix}.$$

We can also transform directly from Alice's coordinates to Claire's coordinates, which is

$$\begin{pmatrix} t'' \\ x'' \end{pmatrix} = \hat{L}(v \oplus v') \begin{pmatrix} t \\ x \end{pmatrix}.$$

and has to be the same. Therefore, the following matrix equation must hold:

$$\hat{L}(v \oplus v') = \hat{L}(v')\hat{L}(v).$$

Performing the multiplication gives us

$$\hat{L}(v')\hat{L}(v) = \gamma_{v'} \begin{pmatrix} 1 & -v' \\ -v' & 1 \end{pmatrix} \gamma_v \begin{pmatrix} 1 & -v \\ -v & 1 \end{pmatrix} = \gamma_{v'}\gamma_v \begin{pmatrix} 1 + v'v & -(v'+v) \\ -(v'+v) & 1 + v'v \end{pmatrix}.$$

Hence, we need

$$\gamma_{v\oplus v'} \begin{pmatrix} 1 & -(v \oplus v') \\ -(v \oplus v') & 1 \end{pmatrix} = \gamma_{v'}\gamma_v \begin{pmatrix} 1 + v'v & -(v'+v) \\ -(v'+v) & 1 + v'v \end{pmatrix}.$$

Comparing coefficients yields the relations

$$\begin{aligned} \gamma_{v\oplus v'} &= \gamma_{v'}\gamma_v \cdot (1 + v'v), \\ \gamma_{v\oplus v'} \cdot (v \oplus v') &= \gamma_{v'}\gamma_v \cdot (v' + v). \end{aligned}$$

Dividing the second of these formulas by the first gives us the addition formula:

$$v \oplus v' = \frac{v + v'}{1 + vv'}.$$

As for the second part of the exercise, we have

$$\begin{aligned}
\gamma_v^{-2}\gamma_{v'}^{-2} &= \gamma_{v\oplus v'}^{-2}(1+vv')^2,\\
(1-v^2)(1-v'^2) &= [1-(v\oplus v')^2](1+vv')^2,\\
1-v^2-v'^2+v^2v'^2 &= 1+2vv'+v^2v'^2-(v\oplus v')^2(1+vv')^2,\\
v^2+2vv'+v'^2 &= (v\oplus v')^2(1+vv')^2,\\
\frac{(v+v')^2}{(1+vv')^2} &= (v\oplus v')^2,\\
v\oplus v' &= \frac{v+v'}{1+vv'}.
\end{aligned}$$

Therefore, (12.2) is fulfilled with (8.3).

Solution to Exercise 53: We use $c = 1$ and the abbreviation $\alpha = (1 - u_x v)^{-1}$. With $\gamma_v^{-2} = 1 - v^2$, we have, from (12.6),

$$\begin{aligned}
\left|\boldsymbol{u}'^2\right| &= u_x'^2 + u_y'^2 + u_z'^2\\
&= \alpha^2[(u_x - v)^2 + (1-v^2)(u_y^2 + u_z^2)]\\
&= \alpha^2[(u_x - v)^2 - (1-v^2)u_x^2 + (1-v^2)|\boldsymbol{u}|^2]\\
&= \alpha^2[(1-u_x v)^2 - (1-v^2) + (1-v^2)|\boldsymbol{u}|^2]\\
&= 1 + \alpha^2 \cdot (1-v^2)\cdot(|\boldsymbol{u}|^2 - 1).
\end{aligned}$$

Setting $|\boldsymbol{u}| = 1$ yields $\left|\boldsymbol{u}'\right| = 1$. From $\alpha^2 \geq 1$, $1 - v^2 \geq 0$ and $|\boldsymbol{u}|^2 \leq 1$, it furthermore follows that $\left|\boldsymbol{u}'\right|^2 \leq 1$.

Solution to Exercise 56: The rest energy of one particle is mc^2 (remember that the mass is meant to be the mass of the resting particle) and its total energy is approximately $mc^2 + m\boldsymbol{v}^2/2$. Summing over all particles gives us a total energy of $E = N_\mathrm{A} \cdot (mc^2 + \frac{1}{2}m\langle \boldsymbol{v}^2\rangle) = N_\mathrm{A} \cdot (mc^2 + \frac{3}{2}k_\mathrm{B}T)$. At room temperature $T_\mathrm{R} = 300\,\mathrm{K}$, the ratio between the average kinetic energy and the rest energy for the particular case when the particles are helium atoms is

$$\frac{\frac{3}{2}k_\mathrm{B}T}{m_\mathrm{He}c^2} = 1.04 \times 10^{-11}.$$

In other words: for "normal temperatures", the kinetic energy of the Helium atom is completely insignificant in comparison to the rest energy.

Solution to Exercise 59: For Alice, $p = 0$ and $E = E(0)$ holds. According to the mentioned transformation, Bob gets $E' = \gamma_v E(0)$ for the energy of the object. Obviously, $E' = E(v)$ must hold, because, for Bob, the object moves with velocity v. This requires that $E(0) + (\gamma_v - 1)mc^2 = \gamma E(0)$ and is fulfilled only for $E(0) = mc^2$.

References

Alväger+64. T. Alväger et al., Test of the second postulate of special relativity in the GeV region. Phys. Lett. **12**, 260 (1964)

Anderson+98. R. Anderson, I. Vetharaniam, G.E. Stedman, Conventionality of synchronisation, gauge dependence and test theories of relativity. Phys. Rep. **295**, 93 (1998)

Ashby03. N. Ashby, Relativity in the global positioning system. Living Rev. Relativ. **6**, 1 (2003)

Bertozzi64. W. Bertozzi, Speed and kinetic energy of relativistic electrons. Am. J. Phys. **32**(7), 551 (1964)

Brecher77. K. Brecher, Is the speed of light independent of the velocity of the source? Phys. Rev. Lett. **39**, 1051 (1977)

Chou+10. C.W. Chou et al., Optical clocks and relativity. Science **329**, 1630 (2010)

Darrigol05. O. Darrigol, The genesis of the theory of relativity, in *Einstein, 1905–2005*, ed. by T. Damour, O. Darrigol, B. Duplantier, V. Rivasseau. Progress in Mathematical Physics, vol 47 (2005)

deAbreuGuerra15. R. de Abreu, V. Guerra, Speakable and unspeakable in special relativity: time readings and clock rhythms. Electron. J. Theor. Phys. **12**, 183 (2015)

Einstein05a. A. Einstein, Zur Elektrodynamik bewegter Körper. Ann. Phys. **322**, 891 (1905)

Einstein05b. A. Einstein, Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig? Ann. Phys. **323**, 639 (1905)

Evenson+72. K.M. Evenson et al., Speed of light from direct frequency and wavelength measurements of the methane-stabilized laser. Phys. Rev. Lett. **29**, 1346–1349 (1972)

Foucault62. J.L. Foucault, Détermination expérimentale de la vitesse de la lumière: parallaxe du Soleil. C. R. Acad. Sci. **55**, 792 (1862)

FrischSmith63. D.H. Frisch, J.H. Smith, Measurement of the relativistic time dilation using μ-mesons. Am. J. Phys. **31**(5), 342 (1963)

HafeleKeating72a. J. Hafele, R. Keating, Around the world atomic clocks: predicted relativistic time gains. Science **177**, 166 (1972)

HafeleKeating72b. J. Hafele, R. Keating, Around the world atomic clocks: observed relativistic time gains. Science **177**, 168 (1972)

IvesStilwell38. H.E. Ives, G.R. Stilwell, An experimental study of the rate of a moving clock. J. Opt. Soc. Am. **28**, 215 (1938) and J. Opt. Soc. Am. **31**, 369 (1941)

KennedyThorndike32. R.J. Kennedy, E.M. Thorndike, Experimental establishment of the relativity of time. Phys. Rev. **42**, 400 (1932)

Kraus+02. U. Kraus et al., Was Einstein noch nicht sehen konnte. Phys. J. **7**, 1 (2002)

LewisTolman09. G.N. Lewis, R.C. Tolman, The principle of relativity, and non-Newtonian mechanics. Proc. Am. Acad. Arts Sci. **44**(25), 711 (1909)

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5

LiebscherBrosche98. D.-E. Liebscher, P. Brosche, Aberration and relativity. Astron. Nachr. **319**, 309 (1998)

Lipson+. A. Lipson, S.G. Lipson, H. Lipson, Optical Physics. ISBN-13 978-0521493451

MansouriSexl77a. R. Mansouri, R.U. Sexl, A test theory of special relativity: I. Simultaneity and clock synchronization. Gen. Rel. Gravit. **8**, 497 (1977)

MansouriSexl77b. R. Mansouri, R.U. Sexl, A test theory of special relativity: II. First order tests. Gen. Rel. Gravit. **8**, 515 (1977)

MansouriSexl77c. R. Mansouri, R.U. Sexl, A test theory of special relativity: III. Second-order tests. Gen. Rel. Gravit. **8**, 809 (1977)

Michelson81. A. Michelson, The relative motion of the earth and the luminiferous ether. Am. J. Sci. **22**, 120 (1881)

MM87. A. Michelson, E. Morley, On the relative motion of the earth and the luminiferous ether. Am. J. Sci. **34**, 333 (1887)

Minkowski08. H. Minkowski, Raum und Zeit, in Jahresberichte der Deutschen Mathematiker-Vereinigung, Leipzig, 1908 (published 1909). For the English translation see H. Minkowski, Space and time, in The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity. ed. by H.A. Lorentz, A. Einstein, H. Minkowski, H. Weyl (Dover, New York, 1952), pp.75–91

Okun1989. L.B. Okun, The concept of mass. Phys. Today **43**, 31 (1989)

Penrose59. R. Penrose, The apparent shape of a relativistically moving sphere. Proc. Cambridge Philos. Soc. **55**, 137 (1959)

Reichenbach58. H. Reichenbach, The Philosophy of Space & Time (Dover, New York, 1958)

Riehle04. F. Riehle, Frequency Standards: Basics and Applications. ISBN 3-527-40230-6

Rizzi+08. G. Rizzi, M.L. Ruggiero, A. Serafini, Synchronization gauges and the principles of special relativity. Found. Phys. **34**, 1835 (2004)

Robertson49. H.P. Robertson, Postulate versus observation in the special theory of relativity. Rev. Mod. Phys. **21**, 378 (1949)

Selleri94. F. Selleri, Theories equivalent to special relativity, in Frontiers of Fundamental Physics, ed. by M. Barone, F. Selleri (Plenum, New York, 1994), p. 181

Terrell59. J. Terrell, Invisibility of the Lorentz contraction. Phys. Rev. **116**, 1041 (1959)

Index

T. Strohm, *Special Relativity for the Enthusiast*,
https://doi.org/10.1007/978-3-031-21924-5