Contents

1. Electricity

1.1. Electric charge

The **electrostatic force** is one of the four fundamental forces that govern the physical world. Many forces, like friction or the normal force, are actually electrostatic forces in disguise. At fine-grain level, electrostatic forces can be explained in terms of interactions between the elementary constituents of matter, protons and electrons. However, it's entirely possible to approach such forces, at least as a starting point, in terms of forces and fields, in the same way as gravity[1].

First, it is a known empirical fact that rubbing certain objects together render them capable of pushing or pulling other objects. For example, rubbing a plastic stick with a piece of wool cloth makes the stick, for a short period of time, capable of pulling towards itself small bits of paper. Also, by rubbing two plastic sticks with a piece of wool cloth and placing them side by side, the two sticks push each other away. Interestingly, when a plastic stick rubbed with a piece of wool cloth is placed side by side with a glass stick rubbed with a piece of silk cloth, the two are drawn closer to each other.

This pushing and pulling, when understood in terms of forces, are defined as the electrostatic force. This force seems to be "induced" by applying friction between two specific materials. When an object acquires the capability of pushing or pulling other objects, it is said to be **electrically charged**, or just **charged** for short. An object that is not charged is said to be **neutral**.

This charge has to exist in (at least) two flavours, since two rubbed plastic rods push each other away but a rubbed plastic stick and a rubbed glass stick pull each other closer. Actually, no experiment has ever found a third flavour of charge: there is no object that, when charged, is capable of pushing or pulling both rods. That is, it always pushes one and pulls the other.

These two flavours are called **positive charge** and **negative charge**: said names are essentially arbitrary (if, from now on, positive charges were to be defined to be negative charges or vice versa, nothing would change), but they hint at the fact that an equal amount of positive charge and an equal amount of negative charge cancel out, like two numbers equal in magnitude but opposite in sign.

One is to be careful in the definition of a charged object and a neutral object. Indeed, a charged object pulls a neutral object onto itself, but the phenomena could also be interpreted as the neutral object pulling the charged object onto itself. The difference lies in the fact that a charged object both pulls neutral objects and objects having the opposite charge, while pushing objects with the same charge a neutral. On the other hand, a neutral object pulls both charges but does nothing with other neutral objects.

Non-rubbed rods cannot pull bits of paper towards themselves, but rubbed rods can. Also, the rubbed rod and the piece of cloth used to rub it are drawn close to each other. This must mean that both the rod and the cloth, initally neutral, already possessed said charge, but in equal amount of both flavours: after rubbing, an "imbalance" is created, with both objects acquiring charge but of opposite flavours.

As the necessity of friction suggests, electric charge is generated by contact. However, not all objects "retain" charge; some objects, when charged, keep the charge for themselves, others let charge pass through them. The first kind of objects are called **insulators**, the second **conductors**.

Even though the charge model is sufficient to explain electromagnetic force, a deeper look into the fine structure of matter reveals its origin. In particular, the charge of objects is to be found in their atoms: each atom is constituted by even smaller costituens, called **particles**. Said particles come in three flavours: **protons**, **neutrons** and **electrons**. Each atom has a mixture of **protons** and **neutrons** tightly bound together at its center, called the **nucleus**, and the electrons orbit around said nucleus.

[1] Indeed, the atomic model (protons and neutrons in the nucleus with electrons orbiting around) came much further in time than the formal treatment of electrostatic forces.

Each proton, neutron and electron in every atom of the Universe has a precise definite mass and a precise definite amount of charge. This charge, just like mass, is a property intrinsic to the particle, and cannot be altered in any way. Any proton has the exact same amount of charge as any electron, but of distinct kind, while neutrons have charge equal to 0. By convention, protons are assigned the positive charge and electrons the negative charge. The attractive electric force between the positive protons and the negative electrons is what keeps the atom in place.

The amount of charge of a single electron is denoted with $-e$, while the charge of a single proton is denoted by $+e$: the plus and minus sign refer to the fact that said charges are assigned to be the negative and the positive charge, respectively. To refer to the magnitude of the charge without referring to its sign, the symbol e is used. The charge of an atom, denoted with q is therefore:

$$q = e \cdot (\text{number of its protons} - \text{number of its electrons})$$

Since most atoms have the exact same amount of protons and electrons, their net amount of charge is 0, because the positive and negative charges even out.

The charge e is called the **elementary charge**, because electrons and protons are the smallest objects in existence that have an electric charge, and it's not possible to split them apart further[2]. This also means that electric charge is **quantized**, meaning that the charge of a large object (made of two or more atoms) is the sum of the charges of each of its atoms, and is therefore an integer multiple of e.

An atom, to be charged, must therefore have more protons than electrons or more electrons than protons. The protons of an atom is tightly bound to the nucleus, therefore it is almost impossible for an atom to "lose" or "acquire" a proton. It is however much more likely for an atom to "lose" or "acquire" an electron. This is because the electron is loosely bound to the nucleus, and a sufficient force (like the friction between a rod and a piece of cloth) can overcome this force, moving said electrons from an atom to another.

The process of removing an electron from an atom is called **ionization**. Ionization creates two atoms where one is in defect of one or more electrons, and is therefore positively charged, and the other has a surplus of one or more electrons, and is therefore negatively charged. The first is also called **positive ion**, the second **negative ion**. Note that charge is **conserved**: if an atom gains a certain amount of positive charge, another atom has to lose the exact same amount of positive charge and the same holds for negative charge. No charge can be created out of nothing or disappear.

The atomic structure also explains why some materials are conductors and some others are insulators. Insulators are materials whose electrons are very tightly bound to their nucleus, whereas conductors are materials whose electrons are very loosely coupled. Even when a conductor has no charge, its electrons are so loose that they almost "drift around" its volume a single, defined atom to which they belong. This means that when electrons are added or subtracted from a conductor, the electric imbalance spreads out quickly. On the other hand, when a conductor is positively or negatively charged, such charging remains localized.

The charge of an object is measured in **Coulomb** (symbol C); the fundamental charge is equal to:

$$e = 1.60 \times 10^{-19} C$$

The electrostatic force that exists between two charged objects can quantified by a formula, known as **Coulomb's Law**:

[2] Actually, this is only true for electrons. Protons are in turn constituted by even smaller particles, called **quarks**, whose elementary charge result in a net charge of $+e$. In particular, a proton consists of two quarks with charge $+2/3e$, called **up quarks** and one quark with charge $-1/3$, called **down quark**. Treating the protons as indivisible units is still better, however, both because it is simpler and because single, isolated quarks do not exist in nature.

$$F_{1 \text{ on } 2} = F_{2 \text{ on } 1} = \frac{1}{4\pi\varepsilon_0}\frac{|q_1||q_2|}{r^2}$$

Where q_1 and q_2 are the charges of the two particles, r is the distance between them and ε_0 is a constant, known as **permittivity constant**. Its value is:

$$\varepsilon_0 = 8.85 \times 10^{-12} C^2/Nm^2$$

Since 4π is itself constant, the expression $1/4\pi\varepsilon_0$ is itself a constant, called the **electrostatic constant**. Said constant is sometimes written as K, whose value is:

$$K = 8.99 \times 10^9 Nm^2/C^2$$

The electrostatic forces are directed along the line joining the two particles: they are repulsive for two charges of the same sign and attractive for two charges of opposite sign.

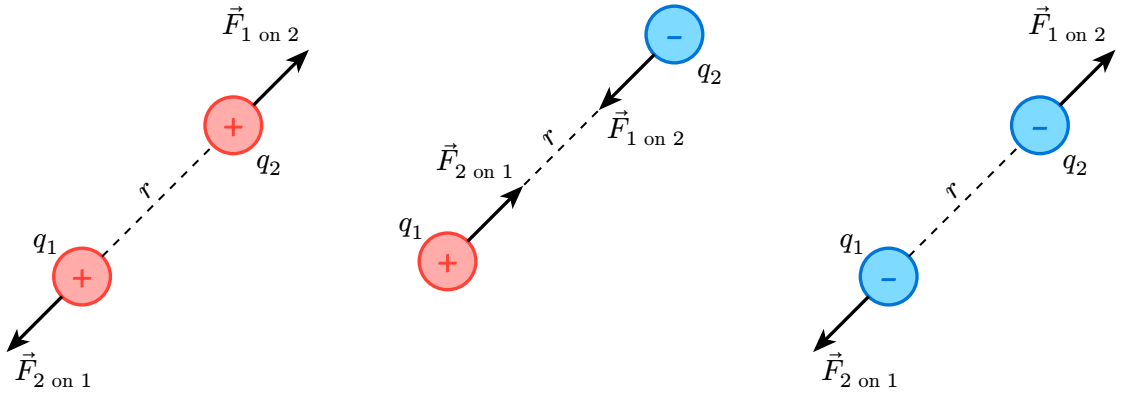


Figure 1: On the left, two positive electric charges push each other away, since their charge is of the same sign. On the right, two negative electric charges push each other away for the same reason. In the middle, a positive and a negative electric charge are drawn closer to each other, since their charge is of the same sign.

Coulomb's Law is very similar to Newton's Law for gravity: both have the two properties of the object at the numerator multiplied together, both are proportional to a constant and both have the distance between the two particles as their denominator. Also, both laws work assuming that the objects are point-like, meaning that their other properties (mass, volume, shape, ecc...) are negligible with respect to their reciprocal distance.

However, while gravity is always an attractive force, electrostatic force exists both in an attractive and in a repulsive fashion. Also, the gravitational constant G is much (much) smaller than the electrostatic constant K, meaning that charges and masses of the same order of magnitude, the electrostatic force is incredibly stronger than gravity.

Exercise 1.1.1: The mass of a proton is $1.67 \times 10^{-27} kg$, while the mass of an electron is $9.11 \times 10^{-31} kg$. What is the ratio between the gravitational force between an electron and a proton and the electrostatic force between an electron and a proton that are $1m$ apart? Assume an electron

Solution:

$$F_e = \frac{1}{4\pi \cdot 8.85 \times 10^{-12} C^2/Nm^2} \left(\frac{\left|1.60 \times 10^{-19} C\right| \cdot \left|1.60 \times 10^{-19} C\right|}{1m^2} \right) = 2.30 \times 10^{-28} N$$

$$F_g = 6.67 \times 10^{-11} m^3 kg^{-1} s^{-2} \left(\frac{1.67 \times 10^{-27} kg \cdot 9.11 \times 10^{-31} kg}{1m^2} \right) = 1.01 \times 10^{-67} N$$

Which means that the electrostatic force between the two is greater than the gravitational force between the two by a factor of:

$$\frac{F_e}{F_g} = \frac{2.30 \times 10^{-28} N}{1.01 \times 10^{-67} N} \approx 2.27 \times 10^{39}$$

□

Electrostatic force is no different than any other force, therefore the superposition principle holds: the net electrostatic force in a system is given by the sum of the electrostatic forces of the single components. Superposition also holds between a mixture of electrostatic and gravitational forces.

Exercise 1.1.2: Three charges q_1, q_2, q_3 are arranged in a triangle at the following coordinates: $(0.0cm, 0.0cm)$, $(5.0cm, 0.0cm)$, $(0, 0cm, 10.0cm)$. Their charges are, respectively: $q_1 = -50nC$, $q_2 = +50nC$, $q_3 = +30nC$. What is the net force acting upon the third charge?

Solution: Since (electrostatic) forces can be superimposed, the resulting force acting upon the third charge is the vector sum between the force coming from the first charge and the force coming from the second charge.

The distance from the second charge to the third charge can be obtained by applying Pythagoras' Theorem:

$$r_{2,3} = \sqrt{r_{1,3}^2 + r_{1,2}^2} = \sqrt{(5.0cm)^2 + (10.0cm)^2} = \sqrt{125.0cm^2} \approx 11.2cm$$

The magnitude of the force acting on the third charge from the first charge is given by:

$$F_{1 \text{ on } 3} = \frac{1}{4\pi \cdot 8.85 \times 10^{-12} C^2/Nm^2} \left(\frac{\left|30 \times 10^{-9} C\right| \cdot \left|-50 \times 10^{-9} C\right|}{0.100m^2} \right) = 1.35 \times 10^{-3} N$$

The magnitude of the force acting on the third charge from the second charge is given by:

$$F_{2 \text{ on } 3} = \frac{1}{4\pi \cdot 8.85 \times 10^{-12} C^2/Nm^2} \left(\frac{\left|30 \times 10^{-9} C\right| \cdot \left|50 \times 10^{-9} C\right|}{0.112m^2} \right) = 1.08 \times 10^{-3} N$$

Consider a system of coordinates centered on the third charge. The force $\vec{F}_{1 \text{ on } 3}$ points downward, hence has no component on the x-axis and its y-component is negative. The force $\vec{F}_{2 \text{ on } 3}$ points north-west, hence has both an x and a y component. The angle θ between $\vec{F}_{2 \text{ on } 3}$ and the coordinate system is:

$$\theta = \arctan\left(\frac{\text{opposite side}}{\text{adjacent side}} \right) = \arctan\left(\frac{10.0cm}{5.0cm} \right) = \arctan(2.0) \approx 63.4°$$

Which means that the two components of $\vec{F}_{2 \text{ on } 3}$ along the x and y axis are:

$$F_{2 \text{ on } 3,x} = \cos(\theta) F_{2,\text{ on } 3} = \cos(63.4°) 1.08 \times 10^{-3} N = 4.83 \times 10^{-4} N$$
$$F_{2 \text{ on } 3,y} = \sin(\theta) F_{2,\text{ on } 3} = \sin(63.4°) 1.08 \times 10^{-3} N = 9.66 \times 10^{-4} N$$

The net force along the x axis is just $F_{2 \text{ on } 3,x}$, since $F_{1 \text{ on } 3}$ has no component in this direction. The net force along the y axis is:

$$F_{3,y} = \left| F_{2 \text{ on } 3,y} - F_{1 \text{ on } 3} \right| = \left| 9.66 \times 10^{-4} N - 1.35 \times 10^{-3} N \right| = 3.84 \times 10^{-4} N$$

Where the minus sign hints at the fact that the two forces point in opposite directions. The magnitude of the resulting force is therefore:

$$F_3 = \sqrt{F_{3,x}^2 + F_{3,y}^2} = \sqrt{\left(4.83 \times 10^{-4} N\right)^2 + \left(3.84 \times 10^{-4} N\right)^2} = 6.20 \times 10^{-4} N$$

The x component of F_3 points leftwards, since $F_{2 \text{ on } 3,x}$ points leftward. The y component of F_3 points downwards, since $F_{1 \text{ on } 3}$ is greater in magnitude than $F_{2 \text{ on } 3,y}$. Therefore, F_3 points south-west. In particular, the angle between F_3 and the x axis is:

$$\phi = \arctan\left(\frac{\text{opposite side}}{\text{adjacent side}}\right) = \arctan\left(\frac{3.84 \times 10^{-4} N}{4.83 \times 10^{-4} N}\right) = \arctan(0.79) \approx 38.5°$$

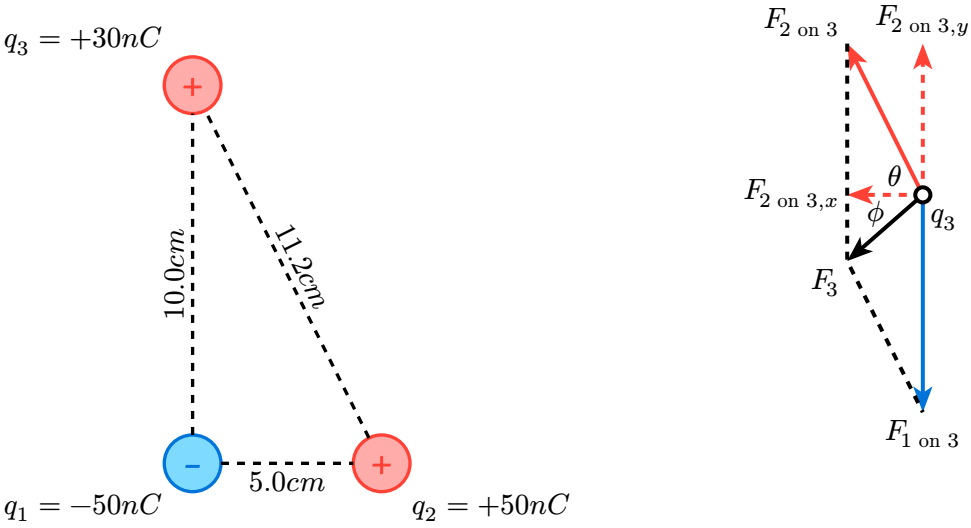


Figure 2: On the left, a graphical representation of the disposition of the three charges. On the right, the forces acting upon the q_3 charge.

□

2. Optics

2.1. Mechanical waves

A **wave** is defined as an organized disturbance moving in space and in time at a given speed. The most known form of wave is the **transverse wave**, where the displacement is *perpendicular* to the direction in which the wave travels. For example, a wave travels along a string in a horizontal direction while the particles that make up the string oscillate vertically.

Most waves necessitate a *medium*, a substance whose particles are displaced from their position of equilibrium when coming into contact with the wave and that return to their position of equilibrium once the wave has passed. These are called **mechanical waves**: water waves, sound waves and the vibrations on the chords of a violin are all examples of mechanical waves. Their speed depend on the medium and on the source that generates it, not on the wave itself.

The waves that are the easiest to treat from a mathematical perspective are the **one-dimensional waves**, waves that move along in a single dimension; in the transverse wave case, this dimension is the y axis. A one-dimensional wave is entirely described by and equation that relates, for all possible points in space and instants in time, its displacement. It is therefore a two-variable functions in the form $D(x,t)$:

$D(x,t)$ = the displacement along the y axis at spacial coordinates x and temporal coordinates t

The most-known form of one-dimensional wave is the **sinusoidal wave**, a wave whose source is a body that is oscillating with **simple harmonic motion** (**SHM**). The equation of a sinusoidal wave at time $t = 0$ is:

$$D(x,0) = A\sin\left(2\pi\frac{x}{\lambda} + \varphi_0\right)$$

A is the **amplitude** of the wave, and represents the highest possible value (in modulo) of displacement that the wave can reach. λ is the **wavelength**, and represents the distance in space between two consecutive amplitudes. The term φ_0 is a **phase constant** that characterizes the initial conditions.

The equation for the displacement of a sinusoidal wave is a periodic function with period λ. This can be shown as:

$$D(x+\lambda,0) = A\sin\left(2\pi\frac{x+\lambda}{\lambda} + \varphi_0\right) = A\sin\left(2\pi\left(\frac{x}{\lambda} + \frac{\cancel{\lambda}}{\cancel{\lambda}}\right) + \varphi_0\right) = A\sin\left(2\pi\frac{x}{\lambda} + 2\pi + \varphi_0\right) =$$
$$= A\sin\left(\left(2\pi\frac{x}{\lambda} + \varphi_0\right) + 2\pi\right) = A\sin\left(2\pi\frac{x}{\lambda} + \varphi_0\right) = D(x,0)$$

The **period** T of a sinusoidal wave is the time needed for the value of the displacement to move between two consecutive amplitudes. The reciprocal of the period is the **frequency** f, that represents the number of times the wave has reached a peak in one unit of time:

$$T = \frac{1}{f}$$

There is an important relationship between the wavelength and the period of a sinusoidal wave: each of its amplitudea travels forward a distance of exactly one wavelength λ during a time interval of exactly one period T. Because speed is distance divided by time, and the speed of a sinusoidal wave is constant, the wave speed must be:

$$v = \frac{\lambda}{T} = \lambda f$$

The velocity of a wave depends on the medium in which the wave is moving: a medium can have more or less inertia and hence be more or less resistant to perturbation. The frequency of a wave depends only on the source that is generating the wave, on the number of "pulses" or "beats" that it generate every unit of time. Hence, the wavelength of a wave depends both on the medium and on the source.

To extend the equation to time istants different from $t = 0$, it is sufficient to point out the fact that $D(x, t)$ is exactly $D(x - vt, 0)$. This is because the sinusoidal wave is periodic with period λ: if the wave had displacement $D(x, t)$ at point x and time t, then it had the exact same displacement at position $x - vt$ when it started moving. Therefore:

$$D(x,t) = D(x - vt, 0) = A\sin\left(2\pi\frac{x - vt}{\lambda} + \varphi_0\right) = A\sin\left(2\pi\left(\frac{x}{\lambda} - \frac{vt}{\lambda}\right) + \varphi_0\right)$$

Since $v/\lambda = 1/T$:

$$D(x,t) = A\sin\left(2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right) + \varphi_0\right)$$

This equation is not only periodic in space with period λ, but it is also periodic in time with period T (hence the name period).

It's possible to write in a more compact form of the wave equation for sinusoidal waves by introducing two auxiliary quantities: the **angular frequency** w and the **wave number** k. The former is 2π multiplied by the frequency, the latter is 2π over the wavelength:

$$\omega = 2\pi f = \frac{2\pi}{T}[s^{-1}] \qquad\qquad k = \frac{2\pi}{\lambda}[m^{-1}]$$

Substituting these quantities in the sinusoidal wave equation gives:

$$D(x,t) = A\sin\left(2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right) + \varphi_0\right) = A\sin\left(\frac{2\pi}{\lambda}x - \frac{2\pi}{T}t + \varphi_0\right) = A\sin(kx - \omega t + \varphi_0)$$

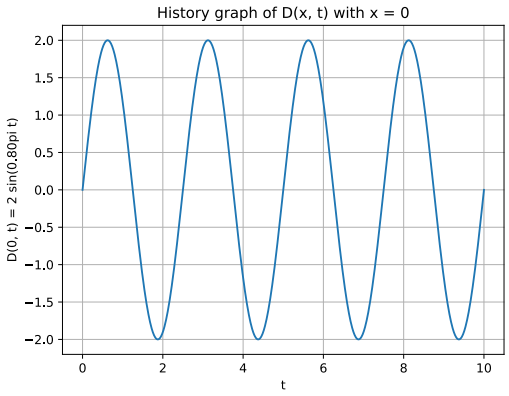
This is the most widely employed form of the equation.

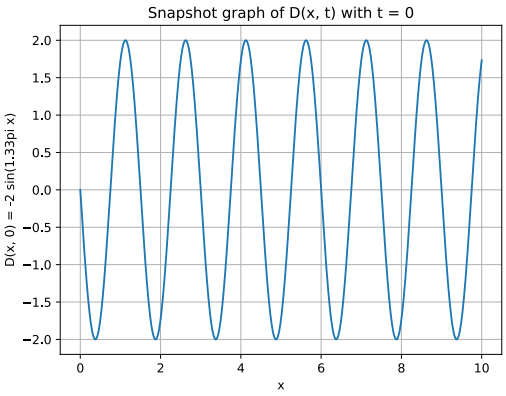
It is often useful to depict the displacement of sinusoidal waves (and of waves in general) graphically, but it presents a challenge. This is due to the fact that the displacement depends both on space and on time, hence it is not possible to represent both at the same time. The only way forward is to set either space or time at a fixed value so that the displacement becomes a single-valued function.

By fixing the spacial coordinate x to a value x_0 and having the temporal coordinate t move freely one obtains the **history graph** of the wave; by fixing the temporal coordinate t to a value t_0 and having the spacial coordinate x move freely one obtains the **snapshot graph**. The history graph represents the evolution of the displacement of the wave in a single point in space with respect to time; the snapshot graph represents the evolution of the displacement of a wave in a single time frame with respect to space.

Exercise 2.1.1: Consider a sinusoidal wave with $A = 2m$, $\lambda = 1.5m$, $\varphi_0 = \pi$ and $T = 2.5s$. What would be its history graph at $x = 0$ and snapshot graph at $t = 0$?

Solution:





□

2.2. Electromagnetic waves and light

Electromagnetic waves are an atypical kind of wave, since they are the only kind of wave that require no medium to be propagated[3]: their "waving" are the self-sustaining oscillations of an **electromagnetic field**. That is, the displacement D *is* an electric or magnetic field. This is why, for example, light (which is a kind of electromagnetic wave) can reach the Earth from the Sun travelling through nothing but empty space.

It has been predicted theoretically and verified experimentally that all electromagnetic waves travel through vacuum with the same speed, called the **speed of light**. The current accepted value of the speed of light, denoted as c, is:

$$c = 299792458 m/s \approx 3 \times 10^8 m/s$$

Electromagnetic waves that the human eye can be perceive fall under the umbrella of **visible light** (or just "light" for short). Visible light encompasses the electromagnetic waves having wavelength between $700nm$ to $400nm$; different wavelength in the range are perceived as different colors. However, there are many more electromagnetic waves other than visible light, all moving at the speed of light and detectable by physical apparatuses. These are classified in the following categories:

- **Radio waves**: from $1 \times 10^4 m$ to $0.1m$, generated by charges accelerated in conducting wires (LC circuits) and are used mostly in communication;
- **Microwaves**: from $0.3m$ to $1 \times 10^{-4} m$, used in radar, to study atomic structures, and to cook;
- **Infrared waves**: from $1 \times 10^{-3} m$ to $7 \times 10^{-7} m$, produced by molecules and objects at room temperature, absorbed by most materials;
- **Visible light**: from $7 \times 10^{-7} m$ to $4 \times 10^{-7} m$, part of the electromagnetic spectrum that our eyes can detect;
- **Ultraviolet waves**: from $4 \times 10^{-7} m$ to $6 \times 10^{-10} m$
- **X-rays**: from $1 \times 10^{-8} m$ to $1 \times 10^{-12} m$, emitted by deceleration of high energetic electrons, and electron transitions; in atoms
- **Gamma rays**: from $1 \times 10^{-10} m$ to $1 \times 10^{-14} m$ emitted mostly by radioactive decays in unstable nuclei.

The arrangement of all possible frequencies and wavelengths that electromagnetic waves can have is called the **electromagnetic spectrum**. Note that the distinction between each class of electromagnetic wave is purely nominal, since there are no fixed boundaries between one and the other.

[3] As touched upon later, matter can also exhibit wave-like properties without the need for a medium. Matter waves are treated separately from classical waves, however, since they don't obey the same laws.

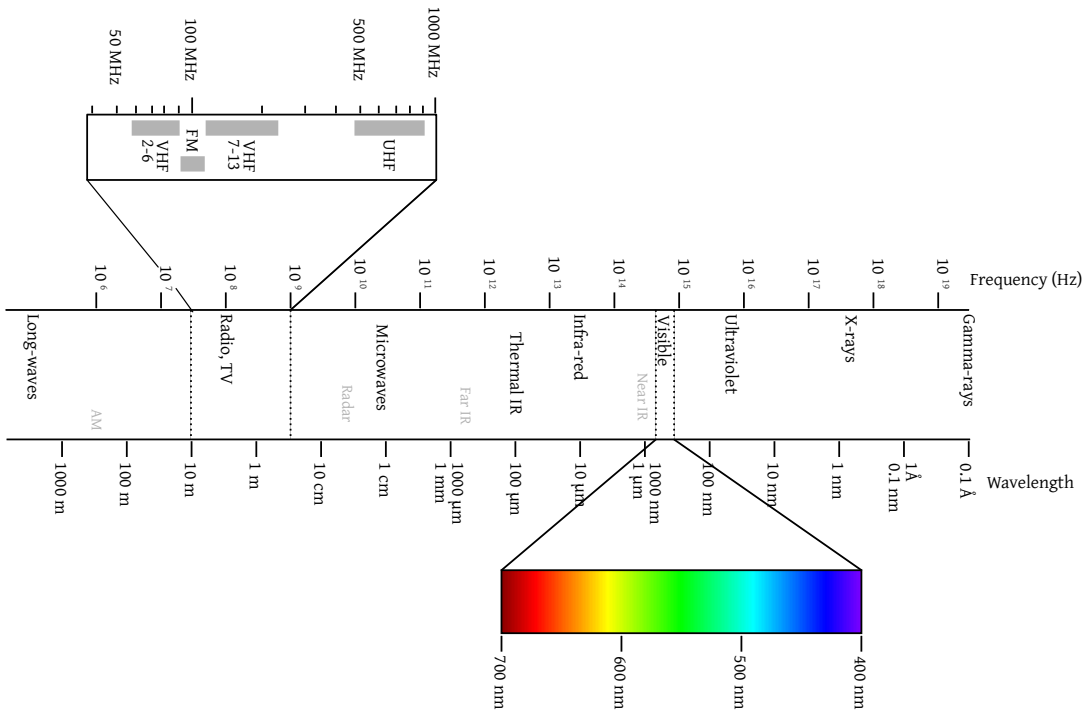


Figure 3: The electromagnetic spectrum. [Original image by Victor Blacus, licensed under the CC BY-SA 3.0 license, based on earlier work by Penubag]

Light is, however, even more elusive than the other electromagnetic waves. This is because treating light merely as a wave is reductive: under different circumstances, different properties of light emerge. For this reason, it is important to distinguish between different models of light, each better suited for describing different phenomena:

- **Light as a wave**. As already stated, light is an electromagnetic wave that travels in a vacuum at constant speed c. Under many circumstances, light behaves no differently than sound or water waves, exhibiting superposition and diffraction. Lasers and electro-optical devices are best described by interpreting light as being a wave;
- **Light as a ray**. When the wavelengths at play are quite long, light travels in bundles of straight lines, as it were a bundle of rays pointing in the same direction. The properties of mirrors and lenses, such as reflection and refraction, are best understood in terms of light rays. The wave interpretation and the ray interpretation are, for the most part, mutually exclusive;
- **Light as a particle**. Light can also be understood as a flow of quantum objects called **photons**. Photons, possessing both wave-like and particle-like properties, arise in treating light from a quantum mechanical perspective, such as interpreting the photoelectric effect or the black-body radiation.

The basis of the ray model of light is the observation that, in everyday life experience, light travels in straight lines, or *rays*, that bounce and/or traverse objects that they encounter.

The ray model is an oversimplification, whose range of validity is confined to the cases where light traverses apertures (lenses, mirrors, and holes) that are very large compared to its wavelength. In this case, light (rays) traverses the aperture without disturbance; if the aperture is shrunk too much, the passage of light would be distorted beyond the capabilities of the ray model, and phenomena such as *diffraction* would emerge.

A **light ray** is defined as an abstract line (does not represent any actual, phyisical quantity) emitted from a *source*, moving in the same direction as the electromagnetic field of light. Any narrow beam of light, no matter how narrow, is actually a bundle of parallel light rays close together. **Lasers**, even though still constituted by many parallel light rays, is as good as an approximation can be of a single, isolated light ray.

Light rays are represented graphically as, as said, straight lines. However, the light as a ray model presupposes that light beams are constituted by an infinite number of rays, hence it is not possible to draw them all. The idea is to draw only a handful of lines, each consistent with the way the rays are spreading, restricting the focus to the lines that are actually worth taking into consideration.

The two main ways light rays are drawn is as rays emitted by a source, going in every direction, or as a bundle of parallel lines moving in the same direction. Source-like points make sense when dealing with objects that are close-by, not only actual light sources but also surfaces that deflect the incoming light to the surroundings. Parallel rays make sense when dealing with very precise sources like lasers or when dealing with far away objects, such as stars, whose incoming light is so distant that the rays arriving at an observer are essentially parallel to each other.

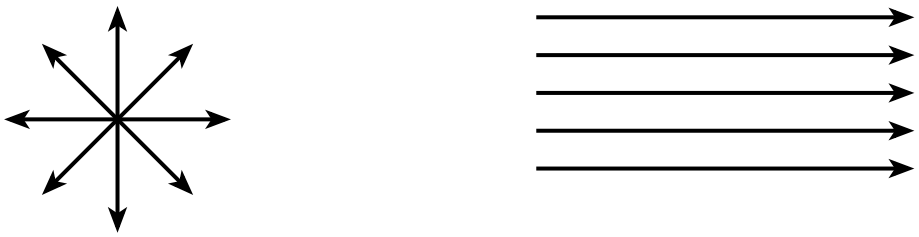


Figure 4: A point-like source emitting rays in every direction (left) and a beam-like bundle of parallel rays (right)

2.3. Reflection and refraction

2.3.1. Reflection

The first phenomena that the ray model of light is better-suited to interpret is **reflection**. Reflection happens when light "bounces back" from surfaces that it comes into contact with. The image of oneself in front of a mirror and the image of the sky mirrored in the water of a pond are examples of reflection. The act of "seeing" itself requires reflection: an observed object is actually the light that it reflected coming into contact with the human eye.

Reflection from a flat, smooth surface, such as a mirror or a piece of polished metal, is called **specular reflection**. It can be observed experimentally that any ray of light hitting a surface that induces specular reflection forms an angle with the normal of the surface (the perpendicular axis drawn in its middle), called **angle of incidence**, that is congruent to the angle formed by the reflected ray with the normal, called **angle of reflection**. It can also be shown that the incoming (incident) ray and the outgoing (reflected) ray are equiplanar. This is referred to as the **law of reflection**.

It is customary to graphically represent specular reflection from a surface in a two-dimensional picture, showing a single ray. This is because, even though the number of rays is infinite and each ray hits the reflecting surface in different points across the surface, the angles of incidence and reflection are the same for all rays.

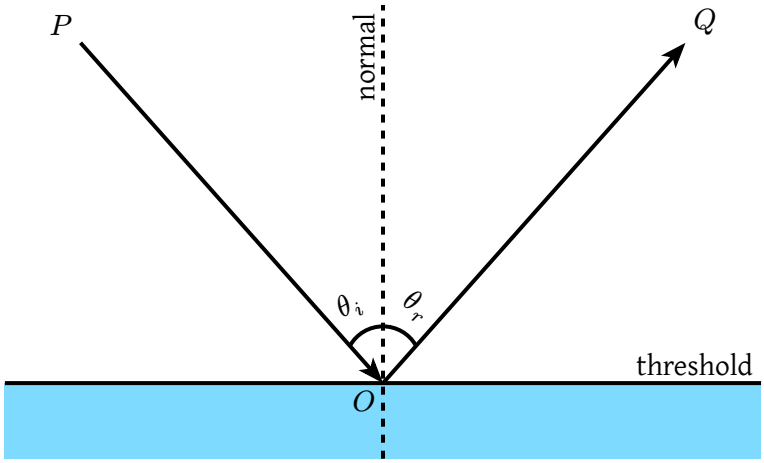


Figure 5: A graphical representation of reflection. θ_i and θ_r are the same angle.

Most forms of reflection are not specular reflection, but **diffuse reflection**. Under diffuse reflection, the law of reflection still holds, but each incident light ray (even though all parallel to each other) is reflected with different angle. Diffuse reflection happens when the surface that is hit is "rough", meaning that it presents irregularities whose size is comparable to the wavelength of incident light. Diffuse reflection is the process that allows most real-world objects to be visually perceived.

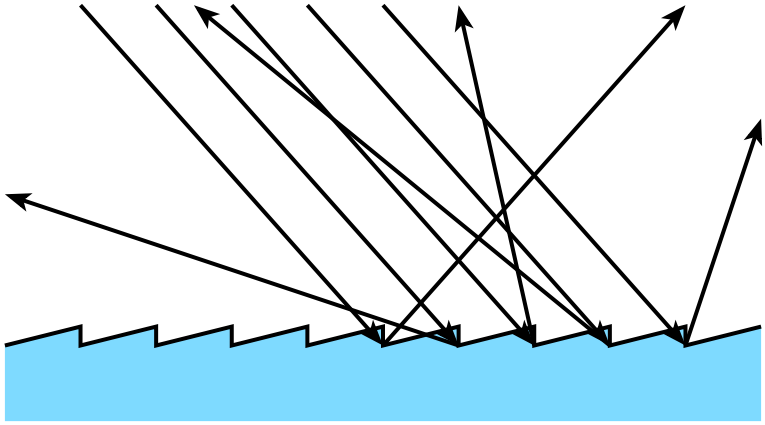


Figure 6: Diffuse reflection on a bumpy surface.

2.3.2. Refraction

Light waves travel with speed c in a vacuum, but they slow down as they pass through transparent materials such as water or glass. The slowdown is a consequence of interactions between the electromagnetic field of the wave and the electrons in the material. The speed of light waves traversing a transparent material is characterized by the material's **index of refraction** n, defined as:

$$n = \frac{c}{v} = \frac{\text{speed of light in a vacuum}}{\text{speed of light in the material}}$$

Clearly, denser materials will have an higher index of refraction, since there are more electrons with which the light wave will interact, lowering its speed. Also, the speed of light in a vacuum is as fast as light waves can go, therefore $n \geq 1$; in the one and only case in which light is moving through a vacuum, $n = 1$.

Since traversing a material slows down a light wave, and since velocity is wavelength times frequency, one of the two (or both) has to change as well. For mechanical waves, the frequency is only dependent on the source, and the same holds for electromagnetic waves; the frequency does not change as the wave moves from one material to another. This means that only the wavelength changes.

In particular, consider a light wave that moves from a vacuum to a material with index of refraction n. In the vacuum, its speed is $c = \lambda f$; in the material, its speed is $v = \lambda' f'$. Being the two frequencies the same:

$$v = \lambda' f' \Rightarrow \frac{c}{n} = \lambda' f' \Rightarrow \frac{\not{c}}{n} = \lambda' \frac{\not{c}}{\lambda} \Rightarrow \frac{1}{n} = \frac{\lambda'}{\lambda} \Rightarrow n = \frac{\lambda}{\lambda'}$$

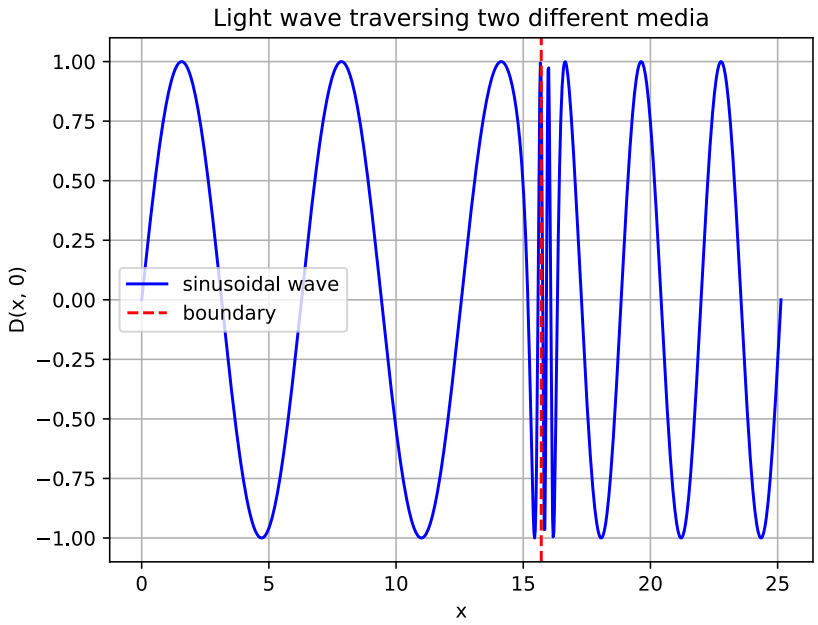


Figure 7: When a light wave moves from a media to another, its frequency stays the same, but its wavelength (and speed) changes.

Another phenomena that is understood in terms of the light as a ray model is **refraction**. Refraction happens when light, or part of the light, traverses the source instead of bouncing back, meaning that is transitions from travelling through a medium to a different medium[4]. In general, refraction happens in tandem with reflection.

When light undergoes refraction, its traversing rays form an angle with the normal with the surface called **angle of refraction**. The angle of incidence and the angle of refraction are not equal, as the direction of the incident light ray and the refracted light ray are not equal. Again, all rays are refracted on the surface at different points but with the same angles and the same directions, hence it can be described by taking into account a single ray.

[4]Note that light is always characterized by the oscillations of an electromagnetic field, not the traversed medium.

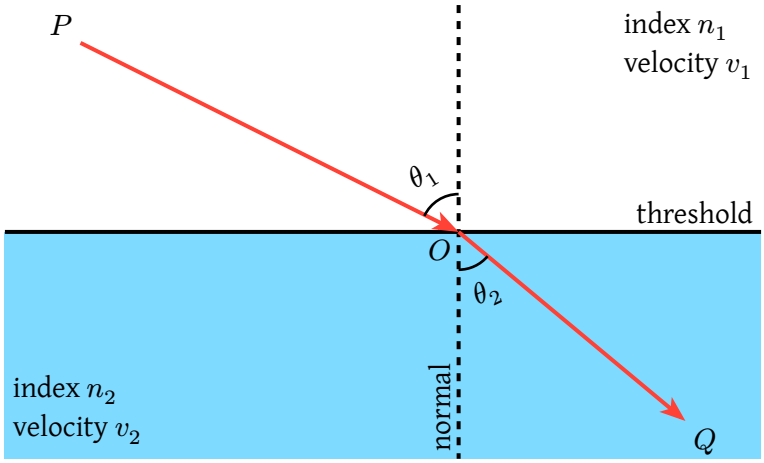


Figure 8: A graphical representation of refraction. θ_1 and θ_2 are different angles.

Consider light rays moving from a medium A with index of refraction n to a medium B with index of refraction n', being refracted. The direction of the light rays (if they move from medium A to medium B or from B to A), are irrelevant. Light will form two angles, the angle of incidence θ and the angle of refraction θ'. These two angles are related by the following empirical law, called **Snell's law**:

$$n \sin(\theta) = n' \sin(\theta')$$

From Snell's law, the following conclusion can be drawn:

- When a ray is transmitted into a material with a higher index of refraction, it bends *towards* the normal;
- When a ray is transmitted into a material with a lower index of refraction, it bends *away* from the normal.

2.3.3. Applying reflection and refraction: measuring the speed of light

Reflection and refraction played a key role in the first (successful) attempt to measure the speed of light from Earth[5]. This is also referred to as **Fizeau's experiment**, bearing the name of its author.

The experimental apparatus consists of a light source, whose light rays reach a glass surface set at $45°$ that reflects and refracts in roughly equal amount. The reflected portion of light is ignored, whereas the refracted portion encounters a spinning toothed wheel: each tooth and each gap are equally spaced, and speed of the wheel can be controlled by the experimenter. If light hits one of the tooth gets blocked, if it hits one of the gaps it passes through. A mirror then reflects the light back to the original glass surface, encountering the wheel once again on its path, being refracted and reflected. The refracted portion of light is discarded, the reflected portion is observed.

[5]Earlier attempts, dating two centuries back, used the rotation period of celestial bodies.

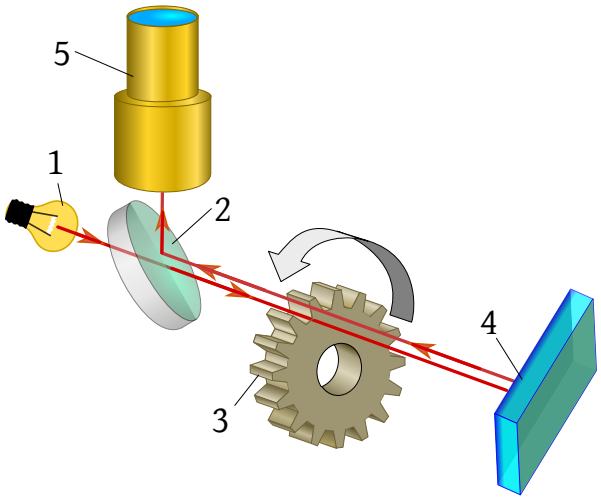


Figure 9: A graphical representation of Fizeau's experimental apparatus: a light source (1) emitting light that is refracted by a glass surface (2), moving over a spinning wheel (3) to a mirror (4) and then back to the glass surface, whose reflection of light is collected (5). [Original image by Д.Ильин, licensed under the CC0 license, based on earlier work by Brews_ohare]

In the starting condition, the wheel is at rest and the light source is aligned with one of the gaps of the wheel; after coming back from the mirror, it will be detected. The speed of the wheel is then increased over and over: the reflected light will become dimmer and dimmer, because more and more light rays will hit the teeth of the wheel when coming back from the mirror.

After reaching a sufficient speed, there will be no incoming reflected light at all, since the wheel and the light will be perfectly synchronized: the light will come back from the mirror to the wheel in the exact same time as a teeth of the wheel will replace the gap that the light used to pass through. If speed is increased even further, the light and the wheel will become once again desynchronized, and the reflected light will become visible once again. When no reflected light can be detected, it must mean that the time that light takes for a round trip (from the wheel to the mirror and back) is the same as the time the wheel takes to rotate for the length of a tooth.

Fizeau's original apparatus had a wheel of 720 teeth, and the distance between the wheel and the mirror was $8633m$. Fizeau observed that reflected light disappeared when the wheel was spinning at a frequency of $12.6s^{-1}$. If the wheel rotates for the exact length of a tooth, the rotated angle is of $\pi/720$. Since $\theta = 2\pi ft$, the time taken for the rotation has to be:

$$t = \frac{\theta}{2\pi f} = \frac{\frac{\cancel{\pi}}{720}}{2 \cdot \cancel{\pi} \cdot 12.6s^{-1}} = \frac{1}{18114.0s^{-1}} = 5.5 \times 10^{-5}s$$

Light has to travel from the wheel to the mirror and back, so the length of the whole path is twice the distance between the mirror and the wheel. Being velocity equal to distance over time:

$$\tilde{c} = \frac{d}{t} = \frac{2 \cdot 8633m}{5.5 \times 10^{-5}s} \approx 313357531.7m/s$$

Which, considering the current accepted value of the speed of light, is off only by:

$$\frac{\tilde{c} \cdot 100}{c} - 100\% = \frac{313357531.7\cancel{m/s} \cdot 100}{299792458\cancel{m/s}} - 100\% \approx 4.5\%$$

2.3.4. Fermat's principle

Even though Snell's law was devised originally as an empirical law, it can be derived by assuming an even more fundamental principle: the **Fermat Principle**. This principle states that light rays, in any circumstance, always travel along the path that requires the least time to be traversed.

To derive Snell's law, consider a light ray that moves from a medium to another. Let A be the starting point (lying in the first medium) and let B be the arrival point (lying in the second medium). Somewhere along the space between A and B, the ray will reach the boundary between the two mediums: let O be the crossing point. Assuming the validity of Fermat's Principle, out of all possible O crossing points the one of interest is the one that minimizes the time needed for the ray to go from A to O and from O to B.

Let H and K be, respectively, the projections of A and B on the boundary between the two media. Since the speed of light in a medium is constant, let $v_1 = c/n_1$ be the speed of the light ray when traversing the first medium and let $v_2 = c/n_2$ be the speed of the light ray when traversing the second medium. Let θ_1 and θ_2 be, respectively, the angle of incidence and of refraction.

Since the spacial coordinates of A and B are fixed at (x_A, y_A) and (x_B, y_B) respectively, the length of the segment $\overline{HK}$ drawn along the boundary is known: let such length be L. The choice of the O point depends entirely on the choice of the length of the segments $\underline{OH}$ and $\underline{OK}$; since the sum of their length is known to be L and the length of one is L minus the length of the other, only one of the two length has to be specified. Arbitrarely, let the length of $\underline{OH}$ be the unknown x and let the length of $\underline{OK}$ be equal to $L - x$.

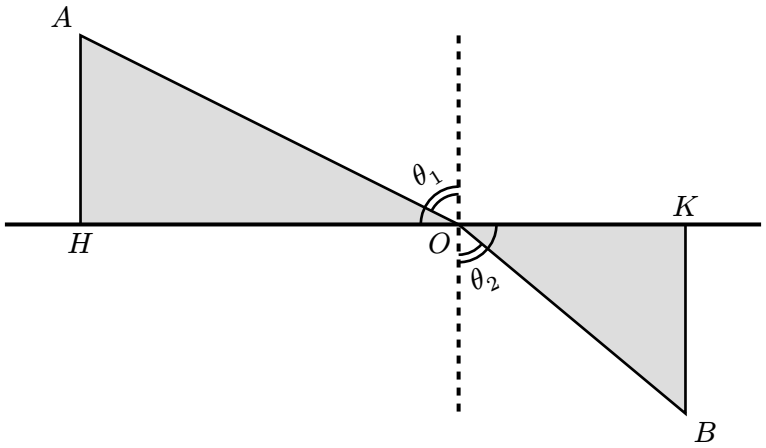


Figure 10: The geometrical setup for deriving the Snell's law.

The segments $\overline{OA}$ and $\overline{OB}$ play the role of the hypothenuse for the right angle triangles AOH and BOK, respectively. This means that:

$$\left\|\overline{OA}\right\| = \sqrt{\|\underline{AH}\|^2 + \|\underline{OH}\|^2} = \sqrt{y_A^2 + x^2} \qquad \left\|\overline{OB}\right\| = \sqrt{\|\underline{BK}\|^2 + \|\underline{OK}\|^2} = \sqrt{y_B^2 + (L-x)^2}$$

Let t_1 be the time needed for the light to go from A to O and let t_2 be the time needed for the light to go from O to B. Since time is space over velocity, these times are the ratio between the lengths of $\overline{OA}$ and $\overline{OB}$ respectively and the speed of light in each medium. The total time that the light has to take for going from A to B is $T = t_1 + t_2$. Using the newlyfound expression for these two lengths:

$$T(x) = t_1(x) + t_2(x) = \frac{\sqrt{y_A^2 + x^2}}{v_1} + \frac{\sqrt{y_B^2 + (L-x)^2}}{v_2}$$

The expression is to be minimized. To do so, it is first necessary to compute its derivative (with respect to x):

$$\frac{dT}{dx} = \frac{dT}{dx}\left(\frac{\sqrt{y_A^2 + x^2}}{v_1} + \frac{\sqrt{y_B^2 + (L-x)^2}}{v_2}\right) = \frac{\frac{dT}{dx}\left(\sqrt{y_A^2 + x^2}\right)}{v_1} + \frac{\frac{dT}{dx}\left(\sqrt{y_B^2 + (L-x)^2}\right)}{v_2} =$$

$$= \frac{\frac{1}{2}\left(y_A^2 + x^2\right)^{-\frac{1}{2}}}{v_1}\left(\frac{dT}{dx}\left(y_A^2 + x^2\right)\right) + \frac{\frac{1}{2}\left(y_B^2 + (L-x)^2\right)^{-\frac{1}{2}}}{v_2}\left(\frac{dT}{dx}\left(y_B^2 + (L-x)^2\right)\right) =$$

$$= \frac{\frac{dT}{dx}\left(y_A^2 + x^2\right)}{2v_1\sqrt{y_A^2 + x^2}} + \frac{\frac{dT}{dx}\left(y_B^2 + (L-x)^2\right)}{2v_2\sqrt{y_B^2 + (L-x)^2}} = \frac{\frac{dT}{dx}\left(y_A^2\right) + \frac{dT}{dx}\left(x^2\right)}{2v_1\sqrt{y_A^2 + x^2}} + \frac{\frac{dT}{dx}\left(y_B^2\right) + \frac{dT}{dx}\left((L-x)^2\right)}{2v_2\sqrt{y_B^2 + (L-x)^2}} =$$

$$= \frac{\cancel{2}x}{\cancel{2}v_1\sqrt{y_A^2 + x^2}} + \frac{\cancel{2}(L-x)\frac{dT}{dx}(L-x)}{\cancel{2}v_2\sqrt{y_B^2 + (L-x)^2}} = \frac{x}{v_1\sqrt{y_A^2 + x^2}} - \frac{L-x}{v_2\sqrt{y_B^2 + (L-x)^2}}$$

Setting the expression equal to 0:

$$\frac{x}{v_1\sqrt{y_A^2 + x^2}} - \frac{L-x}{v_2\sqrt{y_B^2 + (L-x)^2}} = 0 \Rightarrow \frac{x}{v_1\sqrt{y_A^2 + x^2}} = \frac{L-x}{v_2\sqrt{y_B^2 + (L-x)^2}}$$

Multiplying both sides by c:

$$\frac{c}{v_1}\frac{x}{\sqrt{y_A^2 + x^2}} = \frac{c}{v_2}\frac{L-x}{\sqrt{y_B^2 + (L-x)^2}} \Rightarrow \frac{n_1 x}{\sqrt{y_A^2 + x^2}} = \frac{n_2(L-x)}{\sqrt{y_B^2 + (L-x)^2}}$$

Notice how the expressions $x/\sqrt{y_A^2 + x^2}$ and $L - x/\sqrt{y_B^2 + (L-x)^2}$ are the ratio between the adjacent side and the hypothenuse of, respectively, the triangles AOH and BOK. These correspond to the cosines of the angles $\frac{\pi}{2} - \theta_1$ and $\frac{\pi}{2} - \theta_2$, where the $\pi/2$ shift denotes the fact that the angles under consideration are the ones *complementary* to the incidence and refraction angles, not the angles themselves. Which gives:

$$n_1 \underbrace{\left(\frac{x}{\sqrt{y_A^2 + x^2}}\right)}_{\cos(\frac{\pi}{2} - \theta_1)} = n_2 \underbrace{\left(\frac{(L-x)}{\sqrt{y_B^2 + (L-x)^2}}\right)}_{\cos(\frac{\pi}{2} - \theta_2)} \Rightarrow n_1 \cos\left(\frac{\pi}{2} - \theta_1\right) = n_2 \cos\left(\frac{\pi}{2} - \theta_2\right)$$

But since $\cos(\frac{\pi}{2} - \alpha) = \sin(\alpha)$ for any angle α:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$$

Which is precisely Snell's law.

2.3.5. Total internal reflection

Consider a light beam crossing the boundary between two media with index of refraction n_1 and n_2 respectively. Assuming $n_1 > n_2$, due to Snell's law, the refracted light ray will be bent away from the normal. If one were to increase the angle of incidence, keeping both indices the same, the resulting refracted ray will be bent more and more.

At some **critical angle** θ_c, the refracted ray will be exactly parallel to the boundary between the two media. This angle can be found easily by observing how a refracted light ray parallel to the boundary is, by definition, perpendicular to the normal, that is, when the angle of refraction is $\pi/2$. Substituting $\theta_2 = \pi/2$ in Snell's law:

$$n_1 \sin(\theta_c) = n_2 \sin\left(\frac{\pi}{2}\right) \Rightarrow n_1 \sin(\theta_c) = n_2 \cdot 1 \Rightarrow \sin(\theta_c) = \frac{n_2}{n_1}$$

Any incidence angle that is greater than the critical angle will result in no refracted ray at all, since the incident beam is entirely reflected. light rays that experience no refraction and only reflection due to having an incidence angle greater than the critical angle experience the so-called **total internal reflection**.

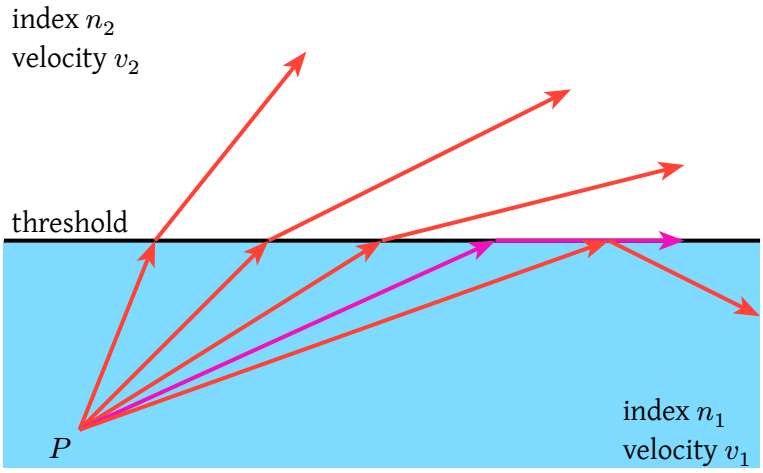


Figure 11: A graphical representation of total internal reflection. The more the angle of reflaction increases, the further the refracted ray is deflected away from the normal. At certain angle, the refracted ray (in purple) is perfectly parallel to the threshold between the two media; any higher angle will result in reflection and no refraction.

The name total internal reflection comes from the fact that, in general, light is never either only refracted or only reflected, but is both. The amount of light that is refracted vanishes as the angle of incidence increases: beyond the critical angle, there is no refraction at all.

Also note that total internal reflection is possible only if light goes from a medium with a greater index of refraction to a medium with a smaller index of refraction. In this case, light will always be partially reflected and partially refracted.

2.3.6. Applying total internal reflection: optic fibers

The most notable application of total internal reflection is light (and information) transmission through **optic fiber** cables. The simplest model of an optic fiber cable is a tube made of glass where light enters from one end and exits from the other end bouncing inside the tube.

This happens because light is reflected along the boundary between the glass of the tube and the outer air. In particular, when light enters the tube, its angle of incidence is specifically tuned to be greater than the critical angle, in order for the light to undergo total internal reflection so that no light is refracted (hence lost). The light rays are below the critical angle (almost perpendicular to the cross section of the tube, actually) when they reach the end of the fiber, thus they refract out without difficulty and can be detected.

This model is far too simplistic for real-world applications, since the air-glass boundary is unreliable: it is still prone to refraction and the tiniest scratch or bruise on the surface of the glass would have it leak light. For this reason, most commercial realizations of an optic fiber cable are constituted by a small-diameter glass tube, called the *core*, nested inside a bigger layer of glass *cladding*. The glasses used for the core and the cladding are specifically designed so that $n_{\text{core}} > n_{\text{cladding}}$, guaranteeing total internal reflection, and to absorb as little light as possible. The boundary between the two is more reliable and not exposed to the environment; to maximize its durability, the cladding is often wrapped inside one or more layers of plastic.

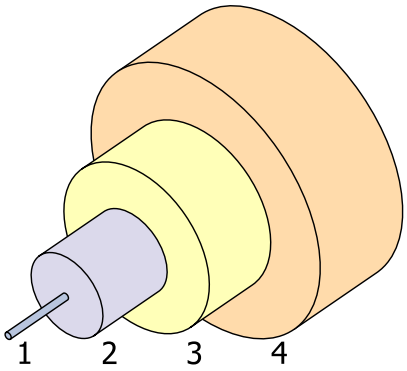


Figure 12: Typical structure of a fiber optic cable, with a core (1), a cladding (2) and two protective plastic caps (3, 4). [Original image by Benchill, licensed under the CC BY-SA 3.0 license, based on earlier work by Bob Mellish]

2.4. Mirrors and lenses

Mirrors are the simplest form of surfaces for which light undergoes specular reflection. The light coming from a self-sustaining light source (a lamp, the Sun, ecc...) hits any object, undergoes diffuse reflection, is reflected from the mirror and reaches an observer.

The image of an object as it appears in a mirror, its **virtual image**, can be thought of as being "on the other side" of the mirror, whose light rays "merge" with the original light rays and moving in a straight line instead of being reflected. Note how the virtual image, as the name suggests, doesn't actually exist: it is not "behind" the mirror, it is just a geometrical construction that explains this apparent optical illusion.

It is a known empirical fact that, for any mirror, the virtual image of an object is specular to the object itself with respect to the mirror. That is, if the object is on the left of an observer, the virtual image is also to the left of the observer (to the right of the mirror) and vice versa.

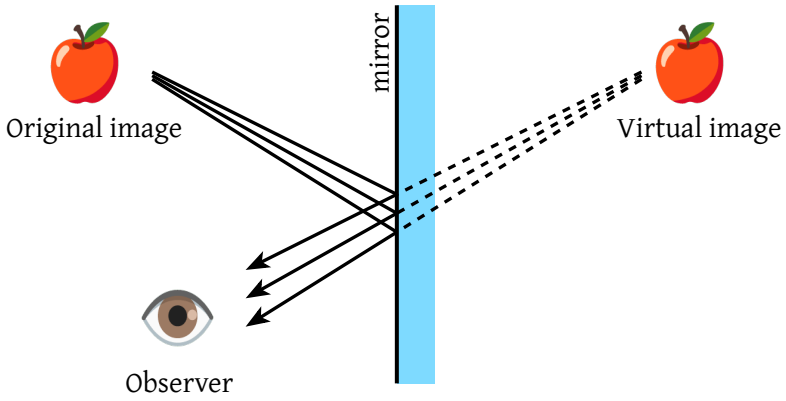


Figure 13: An object in a mirror is reflected as it does because the light rays that it reflects are equivalent to light rays emitted directly from the virtual image. The virtual and original image are specular.

This phenomena can be interpreted assuming Fermat's principle. Not only that, but Fermat's principle also explains why, during reflection, the angles of incidence and of refraction are the same.

Consider a light ray travelling from a point A to a mirror and reflected to a point B. Let B' be the point that is at the same distance from the mirror as it is B. Let C be the intersection between the surface of the mirror and the segment $\overline{AB'}$. Let H be the intersection between the surface of the mirror and the segment $\overline{BB'}$. Let K be the intersection between A and the projection of A on the surface of the mirror.

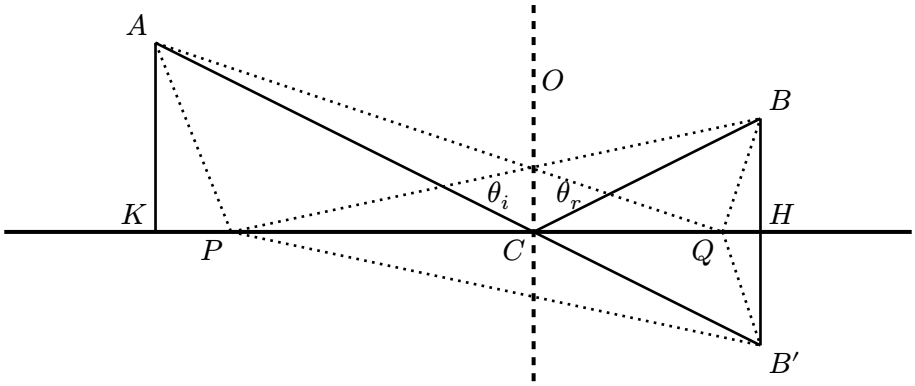


Figure 14: The geometrical setup for deriving the law of reflection. C is the intersection between $\overline{AB'}$ and the mirror, H is the intersection between $\overline{BB'}$ and the mirror. P and Q are alternative candidate points of reflection other than C, chosen at random.

Consider any point P along the surface of the mirror. By construction, the angles $B\hat{H}P$ and $B'\hat{H}P$ are both right angles, and the segments $\overline{BH}$ and $\overline{B'H}$ have the same length. The triangles BPH and $B'PH$ are therefore congruent, since they share one side ($\overline{PH}$), one side is equal to the other and the angle that liest between them is of the same size. In turn, this means that BPB' is an isosceles triangle, therefore $\overline{PB}$ and $\overline{PB'}$ have the same length and the angles $P\hat{B}H$ and $P\widehat{B'}H$ have the same size.

Since, for any point P along the surface of the mirror, $\left\|\overline{PB}\right\| = \left\|\overline{PB'}\right\|$, the total length of the path that goes from A to P to B is the same as the total length of the path that goes from A to P to B'. Of course, the shortest possible path that goes from A to P to B' is the one where light moves in a straight line. That is, the shortest path is the one where the points A, P and B' are aligned, which by construction is the one where $P = C$.

Since light moves at constant speed, the shortest path in terms of time is also the shortest path in terms of distance, and vice versa (note that this does not work with refraction, because light changes speed when going from one medium to another). Therefore, the path of shortest time is the one that goes from A to C to B', which takes just as much as the one that goes from A to C to B.

Since by construction $\overline{AK}$ and $\overline{B'H}$ are parallel (being both perpendicular to the mirror), $C\hat{A}K \cong C\widehat{B'}H$. As already stated, $C\hat{B}H \cong C\widehat{B'}H$, therefore by transitive property $C\hat{B}H \cong C\hat{A}K$. Since by construction $\overline{AK}$, $\overline{BH}$ and $\overline{CO}$ are all parallel to each other, $C\hat{A}K \cong \theta_i$ and $C\hat{B}H \cong \theta_r$. Finally, since $C\hat{B}H \cong C\hat{A}K$, $\theta_i \cong \theta_r$.

Virtual images produced by reflection are (almost) perfect copies of the original, mainatining the same proportions and the same distance from the mirror. Virtual images produced by refraction are not, however: when light rays are refracted, their path is distorted, creating a virtual image that does not coincide with the original image.

Consider an object placed in a medium with index of refraction n_1 and an observer placed in a medium with index of refraction n_2. Suppose, without loss of generality, that $n_1 > n_2$. The observer is able to see the object either because light from a source hits it and undergoes diffuse reflection or because the object is itself a source of light. In any case, light rays move from the object to the observer, but when transitioning from the first medium to the second medium undergoes refraction. The observed object is found by tracing the light rays backwards, but since the incoming rays were deflected away from the normal due to refraction, their point of juncture is not the original object, but a point closer to the threshold. This means that the observer will perceive the object to be closer than it really is.

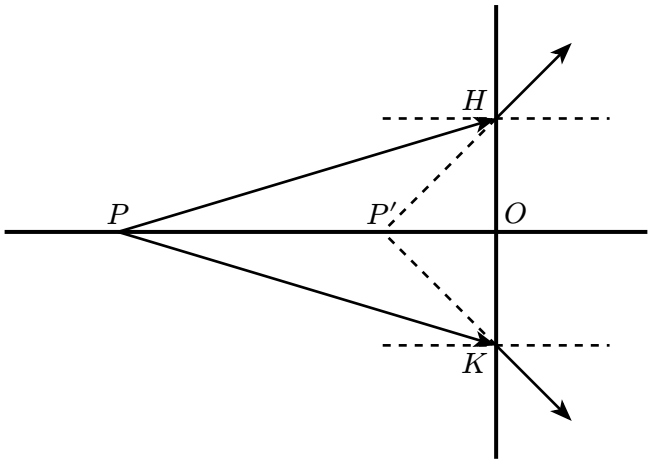


Figure 15:

A **lens** is a transparent object that uses refraction at curved surfaces to form an image from diverging light rays. That is, a lens receives incident light rays as "input" from an object and "outputs" light rays that create a copy of the object. Most lenses fall into two categories: **converging** and **diverging** lenses. Converging lenses bend the light rays towards the optical axis, diverging rays bend light away from the optical axis.

The light rays outgoing from a lens all intersect at a point called the **focal point**. The focal point of a lens depends only and exclusively on the lens itself (on the material that is built of, on its shape, ecc...), not on the incoming light. The focal point of a converging lens is on the other side of the object, it's the point where all outgoing light rays converge (hence the name). The focal point of a diverging lens is on the same side of the object, and represent the mathematical construction that interprets the point where such light rays originate from. The distance of the focal point from the lens is called the **focal length** f of the lens.

3. Theory of relativity

3.1. Introduction

The two most successful physical theories of the 19th century, (classical) mechanics and electromagnetism, stood the test of time and work almost seamlessly hand in hand. There is however one troublesome aspect that, when taken into account, has the two theories completely at odds: light.

Maxwell's Equations for electromagnetism, in accord to the experimental results, predict that light should travel with speed:

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} = \frac{1}{\sqrt{(8.854 \times 10^{-12} F/m)(1.256 \times 10^{-6} N/A^2)}} =$$
$$= \frac{1}{\sqrt{11.120 \times 10^{-18} (s^4 \cdot \cancel{A^2} \cdot \cancel{kg} \cdot \cancel{m} \cdot \cancel{s^{-2}} / \cancel{kg} \cdot m^2 \cdot \cancel{m} \cdot \cancel{A^2})}} =$$
$$= \frac{1}{\sqrt{11.120 \times 10^{-18} s^2/m^2}} = \frac{1}{3.334 \times 10^{-9} s/m} \approx 3.000 \times 10^8 m/s$$

That is, light apparently moves at speed c in any circumstance. This is problematic, because mechanics postulates that there is no such thing as an "absolute velocity": all velocities depend on the reference frame from which they are observed, as the principle of Galilean relativity states.

Consider two inertial reference frames, S and S', where S' is moving at constant speed v (along the x axis) with respect to S. Suppose that a beam of light were to be travelling at speed c (again, along the x axis) with respect to S; then, applying Galilean relativity, one would expect the light beam to travel at velocity $c - v$ with respect to S'. However, when measuring the speed of the light beam with respect to S' one still finds that it moves with speed c.

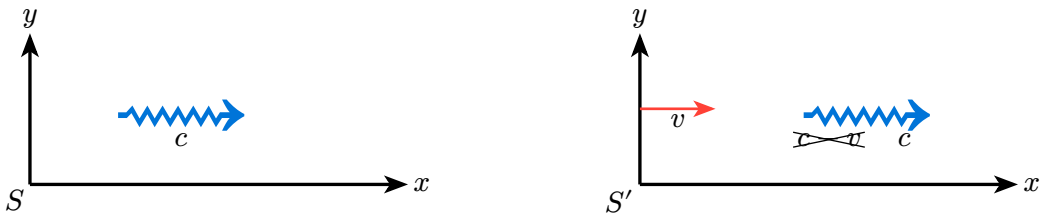


Figure 16: With respect to an inertial frame of reference S, a light beam (in blue) moves with velocity c. With respect to another inertial frame of reference S', that is moving with respect to S at constant velocity v, it would be assumed that the light beam would travel with velocity $v - c$ with respect to S'. However, this is not the case: the light beam moves at velocity c both with respect to S and with respect to S'.

Also, despite its accomplishments, Maxwell Equations left a question opened: how could light travel in the absence of a medium? Light was understood to be a wave, and each wave needed a medium through which they propagate (like air is the medium through which sound propagates, or water is the medium through which water waves propagate). However, how can light travel, for example, from the Sun to the Earth with nothing but empty space in between?

One hypothesis that was advanced early on to reconcile the two theories and, at the same time, to solve this conundrum, was to postulate that the entire Universe is permeated by a peculiar substance called **luminiferous ether**, or just **ether**. This ether would be the medium through which light propagates, and the speed of light is constant when measured with respect to the "special" frame of reference of the ether. Stated otherwise, the value of c that arises from Maxwell equations does not refer to the speed of light "per se", but the value measured with respect to this priviledged frame of reference.

The ether hypothesis had some ground, since Maxwell's Equations invalidated Galilean coordinate transformations. That is, when observing the same experiment involving light from different frames of reference, the results would be different, even if the frames are inertial. This means that it's possible to rely on Maxwell's Equations to tell apart which system is moving with respect to the other in an "absolute" sense, and "absolute" motion is explicitly forbidden in Newtonian/Galilean mechanics.

If this hypothesis were to be true, the contradiction would be solved, even though this would elevate the ether's reference frame as a "special" reference frame, in which motion is absolute. Given an inertial frame of reference S, let v be the velocity with which S moves with respect to the frame of reference of the ether. If it were to be possible to compute v, now the speed of light measured with respect to S would now be $c + v$ (or $c - v$, depending on the direction), as expected.

The ether hypothesis was short-lived, however, since it became clear that no such substance exists. Even before experimental evidence proved the hypothesis wrong, it could hardly be possible for this ether to exist, since it would be something that is present in the entire Universe and yet being barely noticeable.

The most famous experiment that put the existence of the ether into question was the **Michelson and Morley experiment**.

A different approach was the one followed by Einstein, in its **special theory of relativity**: the name "special" comes from the fact that the theory concerns itself only with inertial frames of reference, hence "special" as in "special case". The theory is based on only two postulates:

- **Principle of relativity**. All physical laws are the same with respect to any inertial frame of reference. This is an extension of the Galilean Principle of Relativity, which states that (just) the laws of mechanics are the same with respect to any inertial frame of reference;
- **Constant speed of light**. As predicted by the Maxwell Equations, the speed of light is a universal constant, that holds the same value in any inertial frame of reference.

With these two assumptions, it is possible to construct a theory that is consistent and solves (almost) all contradictions between mechanics and electromagnetism. It introduces, however, many seemingly paradoxical consequences.

First of all, it is necessary to let go of the notion of having a "global" time that flows at the same pace for any reference frame. This is the case in Galilean transformations, where the time variable t in a given (inertial) reference frame does not change when considering the same even in a different (inertial) reference frame.

Consider an experimental setup where two sources of light, one facing the other, are placed vertically at a given distance L. The two sources are kept fixed in place, so that they either both stand still or they both have to move at the same speed. Whenever one source captures a photon emitted by the other, it sends a photon back.

If the apparatus is standing still, both from the frame of reference of an external observer and from the frame of reference of the light source, the photon has to travel along a straight line to start from one source, reach the other source and go back to the original source.

If, on the other hand, the apparatus is moving at constant velocity (for simplicity, only along the left-right axis) with respect to an external observer, the two frames of reference paint a different picture. For the frame of reference of the source, the path traced by the photon is still a vertical line, since it's moving along at constant speed. On the other hand, from the frame of reference of an external observer the path traced by the photon is two incident lines, since both the photon *and* the apparatus are moving.

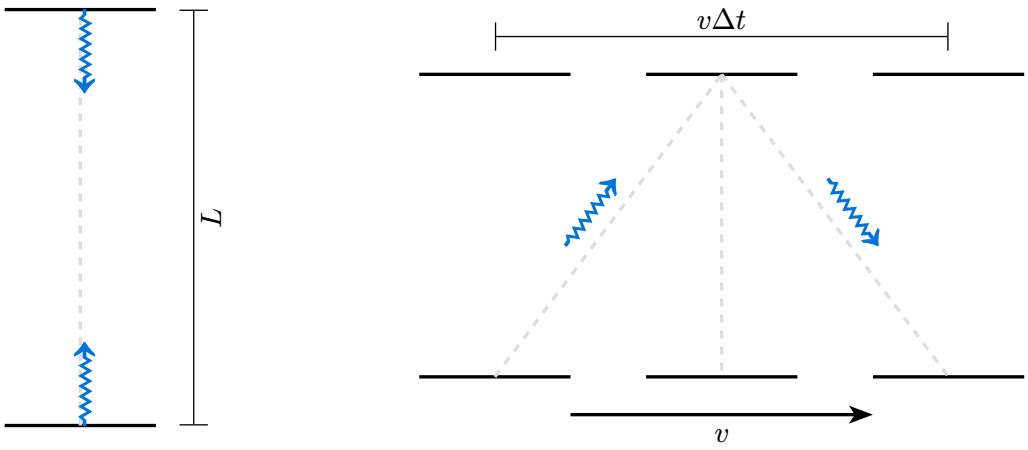


Figure 17: The apparatus consists in two facing sources of light exchanging photons. When the apparatus is still (left), the distance that a photon has to travel to go from its source to the other source and back to its original source is $2L$. When the apparatus is moving (right) the light has to travel a a distance longer than $2L$ to go back and forth, when observed from an external frame of reference.

Let v be the velocity with which the apparatus is moving with respect to an external observer. Let t_0 be the time frame in which the photon starts moving from its original source and let t_f be the time frame in which the photon has returned to its source. Let then $\Delta t = t_f - t_0$ be the time interval in which the photon moves.

In a Newtonian setting, the velocity of the photon when the apparatus is still would be c (as predicted by Maxwell's Equations), both in the frame of reference of the source and in the frame of reference of an external observer. On the other hand, when the apparatus is moving the velocity would be c in in the frame of reference of the source and $c + v$ in the frame of reference of an external observer. Also, the time interval Δt would be the same in both frames of reference and in both situations.

However, if Einstein's First Postulate is to be taken into account, the velocity for both frames of reference when the apparatus is moving has to be c. The distance travelled by the moving apparatus (with respect to an external observer) is $v\Delta t$, the velocity of the apparatus multiplied by the time the photon has taken to go back and forth once. The distance between the two light sources is still L, so the distance d travelled by the photon with respect to an external observer can be computed applying Pythagoras' Theorem:

$$\frac{d}{2} = \sqrt{\left(\frac{v\Delta t}{2}\right)^2 + L^2} \Rightarrow \frac{d^2}{4} = \frac{v^2(\Delta t)^2}{4} + L^2 \Rightarrow d^2 = v^2(\Delta t)^2 + 4L^2$$

With respect to the frame of reference of the source, the photon still travels along a straight line of length L at (constant) velocity c. However, since the velocity of the photon must be c for any observer and any inertial frame of reference, the length L must be equal to $c\Delta t'$, where $\Delta t'$ is a time interval having different size with respect to Δt:

$$d^2 = v^2(\Delta t)^2 + 4L^2 = v^2(\Delta t)^2 + c^2(\Delta t')^2$$

Where the factor of 4 was included directly into the new time $\Delta t'$, being a dummy variable.

Since the distance d must also be equal to $c\Delta t$:

$$c^2(\Delta t)^2 = v^2(\Delta t)^2 + c^2(\Delta t')^2 \Rightarrow c^2(\Delta t')^2 = (c^2 - v^2)(\Delta t)^2$$

Solving for Δt:

$$(\Delta t)^2 = \left(\frac{c^2}{c^2 - v^2}\right)(\Delta t')^2 \Rightarrow \Delta t = \Delta t' \sqrt{\frac{c^2}{c^2 - v^2}} \Rightarrow \Delta t = \Delta t' \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \Rightarrow \Delta t = \gamma \Delta t'$$

Where $\gamma = 1/\sqrt{1 - \frac{v^2}{c^2}}$ is also referred to as **Lorentz's factor**. This result does not hold just for this experimental setup, but for any pair of inertial frames of reference where one is moving at a constant speed with respect to the other.

The equation implicitly states that no velocity can be greater than the speed of light, in any inertial frame of reference. If this wasn't the case, the fraction v^2/c^2 could become greater than 1, potentially resulting in a value of γ (and of $\Delta t'$) that is a complex number. Since time intervals are necessarely real numbers, any velocity must be smaller than c, which also implies that γ is always greater or equal than 1.

This means that time effectively runs faster or slower depending on the frame of reference in which events are observed. In particular, when two inertial frames of referenece are at rest with respect to each other ($v = 0$), then $\gamma = 1$, meaning that time flows at the same pace in both reference frames. On the other hand, as v approaches c, then γ approaches $+\infty$, meaning that as velocity increases as time slows down with respect to a frame of reference of an external observer.

Since the time observed in a frame of reference at rest with respect to the events is the smallest time interval than can be observed, being $\gamma = 1$ the smallest possible value of γ, this time interval is also referred to as **proper time**. It should be noted that all measured times in all inertial frames of reference moving at any constant speed are equally valid observations of the same phenomena. The name "proper time" refers to the fact that, since the observer is itself not moving, there is no "distorsion" induced by velocity when measuring time.

Exercise 3.1.1: What would be an analogy?

Solution: The Moon, when observed from the Earth, appears to be no wider than a couple of centimetres. However, the *real* width of the Moon is roughly $3.5 \times 10^6 m$, which is orders of magnitude larger. This happens because the distance from which a phenomena is observed modifies its apparent proportions: the *real* width of the Moon is the one observed when being close to it, because it's the least "distorting" point of view. This does not mean that the width of the Moon observed by the Earth is *wrong*, just that its point of view is more biased. In the same fashion, proper time describes a phenomena with the least bias, and hence it is understood to be the *real* temporal description. □

The phenomenon of time moving at slower pace in external frames of reference is called **time dilation**. Time dilation is an observable property of time and space, and it has real physical consequences. Any real-world process that can be observed, from supernovas to aging, is subjected to time dilation: a person ages faster or slower based on which frame of reference they are observed, for example.

Exercise 3.1.2: Suppose that an observer standing still is watching an alien spaceship passing by left-to-right at 85% the speed of light. If the time interval from when the spaceship is first sighted by the observer to when it is right in front of them is, from its reference frame, 2 minutes, how much time is passed in the same time frame in the frame of reference of the spaceship?

Solution: If $v = 85\%c$, then v^2/c^2 is $(0.85)^2 \approx 0.73$, which gives a Lorentz's Factor of $1/\sqrt{1-0.73} \approx 2$. If the time interval in the reference frame of the observer is $2min$, the time interval in the reference frame of the spaceship (the proper time) is $2min/2 = 1min$. □

It should be noted, however, that γ scales incredibly slowly with velocity, and only velocities that are very close to the speed of light can yield a value of γ large enough to measure significative time offsets.

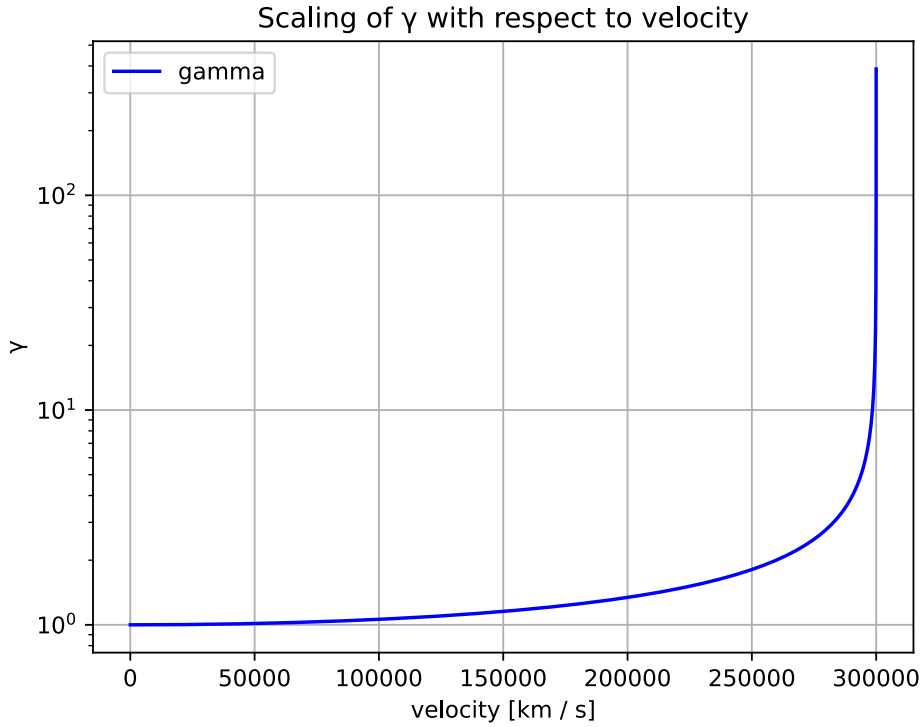


Figure 18: Value of γ with respect to different values of velocity. As can be appreciated, γ becomes noticeable only at incredibly high values of c (γ is displayed in logarithmic scale).

Also, since most real-world phenomena have a velocity that is no where near the speed of light, taking into account the γ factor is not necessary, because $\gamma \approx 1$ and hence $\Delta t' \approx \Delta t$. This is why Newtonian mechanics is still a reasonable model for interpreting reality when working at slow velocities.

Exercise 3.1.3: Fastest human-made space probes have a velocity of roughly $200km/s$. If one such probe were to be observed from an inertial frame of reference, what would be the Lorentz's Factor?

Solution: The speed of the probe is roughly 0.0006 times the speed of light. Hence:

$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} = \frac{1}{\sqrt{1-\left(\frac{6\times10^{-4}\not{c}}{\not{c}}\right)^2}} = \frac{1}{\sqrt{1-3.6\times10^{-8}}} \approx 1.00000018$$

□

The fact that the flow of time varies with different frames of reference also implies that lenghts vary with different frames of reference. That is, the same object can be observed to be longer or shorter depending on which frame of reference it is observed.

Consider, again, two inertial frames of reference, where one is moving with respect to the other in the left-right direction at constant speed v. Suppose that the first measures a certain time interval Δt when moving from point A to point B, travelling a certain distance L. Being a constant uniform motion, $L = v\Delta t'$.

The second frame of reference will instead measure a different time interval Δt when moving from B to A. If the travelled distance were to be the same as the distance observed in the other frame of reference, one would have $L = v\Delta t$, since the velocity is the same (in modulo) for both inertial frames of reference. This would mean $v\Delta t = v\Delta t'$, which entails $\Delta t = \Delta t'$. This is not possible, however, since by construction the two frames of reference are moving with respect to each other.

It is therefore necessary to assume that the second frame of reference actually observed a distance of $L' = v\Delta t$, with L' being different from L. Recalling that $\Delta t' = \gamma \Delta t$:

$$L' = v\Delta t = v\frac{\Delta t'}{\gamma} = \frac{L}{\gamma}$$

This is also referred to as **length contraction**, or **Lorentz's contraction**. Analogously to how proper time was defined, **proper length** is a length measured from the frame of reference of an external observer, which is also the longest possible.

Figure 19: In the frame of reference of an external observer (left), a spaceship moving from the Moon to Saturn would appear "squeezed", but the distance between the planets would be its proper length. In the frame of reference of the spaceship itself (right), the length of the spaceship would be its proper length, but the distance between the planets would appear "squeezed". [Emojis retrieved from Openemoji and licensed under the CC BY-SA 4.0 license]

Note that, this way, special theory of relativity presents a remarkable symmetric: a frame of reference at rest with a moving body will measure proper time but not proper distance, whereas a frame of reference external to the moving body will measure proper distance but not proper time. In particular, in the first frame of reference both time and space are shorter, in the second both time and space al longer.

It should be noted that, even in Special Relativity, some physical quantities are invariant among different frames of reference. First, note that length contraction happens only along the direction of motion, not along all direction. Consider two apparatuses such as the two mirror example, moving with respect to each other and with respect to a third, external frame of reference at velocities v_1 and v_2.

Assume that both apparatuses start in the same condition and in the same spacial coordinates. An observer in the external frame of reference would observe the photon in the first apparatus travel $\frac{1}{2}\Delta x_1 m$ in the x direction and $\frac{1}{2}c\Delta t_1 m$ in the y direction, while observing the photon in the second apparatus travel $\frac{1}{2}\Delta x_2 m$ in the x direction and $\frac{1}{2}c\Delta t_2 m$ in the y direction. In both cases, the height L is the same; applying Pythagoras' Theorem:

$$L^2 = \left(\frac{1}{2}c\Delta t_1\right)^2 - \left(\frac{1}{2}\Delta x_1\right)^2 = \left(\frac{1}{2}c\Delta t_2\right)^2 - \left(\frac{1}{2}\Delta x_2\right)^2$$

Cancelling the $\frac{1}{2}$ factor:

$$c^2\Delta t_1^2 - \Delta x_1^2 = c^2\Delta t_2^2 - \Delta x_2^2$$

The quantity $c^2\Delta t^2 - \Delta x^2$, also called the **spacetime interval**, is the same in any frame of reference.

Special relativity introduces another seemingly counterintuitive result: two events can or cannot appear to be simultaneous depending on the frame of reference in which they are observed. This is because the act of *observing* is itself dependent on the speed of light: be it a measuring instrument or the human eye, *observing* a phenomena entails capturing a light beam carrying the information associated to said phenomena, moving from the phenomena to the observer. Since this movement happens at the speed of light (that is, is not istantaneous), there is necessarely a delay. Also, since the speed of light is always the same in any inertial frame of reference, changing frame of reference also potentially changes the distance that the light has to travel and therefore the perceived time frame at which an event is observed.

Consider two inertial reference frames, one moving with respect to the other at velocity v. Suppose that one of the two frames is centered in the middle of two events, that happen simultaneously. This means that the light wave carrying the information associated to the two events reaches the observer in said frame of reference (moving at speed c) reach the observer in the same time frame. The other reference frame, having to take into account the movement of the frame of reference, will not perceive both events happening simultaneously: one will be perceived before the other.

Figure 20: In both situations, both events (a camera flash) are synchronized so that they happen simultaneously. In the first situation, an external observer lies exactly in the middle of the two, and therefore will always observe the two events simultaneously. In the second situation, the observer is closer to the of the two, and therefore will observe one (the closest) before the other.

Exercise 3.1.4: Muons are subatomic particles created in the upper atmosphere when cosmic rays collide with air molecules. Muons are unstable particles that have an half-life of $1.5\mu s$, meaning that, given a sample of muons, each $1.5\mu s$ roughly half the sample size decays into other smaller particles. Their speed is very close to the speed of light, and their presence on Earth's surface can be easily detected with ad-hoc particle detectors.

The upper atmosphere is about $60km$ above sea level, and the speed of a muon is about $0.9997c$. The time that a muon takes to travel from the atmosphere to the Earth's surface moving at constant speed straight in a downward direction (without decaying) is:

$$\Delta t \approx \frac{h}{c} = \frac{6 \times 10^4 m}{3 \times 10^8 m/s} = 2 \times 10^{-4} s = 200 \mu s$$

Since $200\mu s / 1.5\mu s$ is 133, it means that a muon that were to take such path would have to "dodge" decaying 133 times in a row. Since the probability of decaying (and of not-decaying) is always 0.5, the probability for a muon to reach the Earth without having decayed would be $(0.5)^{133} \approx 10^{-40}$.

This would mean that, every 10^{40} muons, only a single muon would reach the Earth's surface. However, experimental data contradicts such observation, since around 10% of the muons created in the upper atmosphere are detected on Earth. How can this be?

Solution: This apparent paradox can be solved by taking Special Relativity into account. From the muon's frame of reference, it is the Earth that is moving towards it at speed $0.9997c$: if an

observer in the frame of reference of the Earth's surface measures a (dilated) time of $200\mu s$, the muon observes a (proper time) of:

$$\Delta t = \frac{\Delta t'}{\gamma} = 200\mu s\sqrt{1 - \frac{v^2}{c^2}} = 200\mu s\sqrt{1 - (0.9997)^2} = (0.0245 \cdot 200)\mu s = 4.9\mu s$$

Which gives an half-life of $4.9\mu s / 1.5\mu s \approx 3.3$. This means that the "true" probability of a muon to reach the Earth's surface is $(0.5)^{3.3} \approx 0.1$, as confirmed by empirical evidence. □

Galilean transformations allow one to convert a set of coordinates and velocities that are relative to a given frame of reference S to a set of coordinates relative to a different frame of reference S'. There is clearly interest in extending Galilean transformations to account for the phenomena of Special Relativity: such an extension should abide to three constraints:

1. Fall back to Galilean transformations at low speeds;
2. Take into account both transformations in space and transformations in time;
3. Ensure that the speed of light remains the same in all reference frames.

An "educated guess" would be an expression of the form:

$$x' = A(x - vt) \qquad\qquad x = A(x' + vt')$$

Where A is a to-be-determined constant that goes to 1 as the speed v goes to 0. To determine it, assume that both reference frames S and S' start in the same position $x = x' = 0$ and start in the same time frame $t = t' = 0$.

An event happens at coordinates (x, t) with respect to S and at coordinates (x', t') with respect to S'. The information carried by such events travels at the speed of light for a distance of $x = ct$ in S and $x' = ct'$ in S'. Substituting in the previous expressions:

$$ct' = A(ct - vt) = At(c - v) \qquad\qquad ct = A(ct' + vt') = At'(c + v)$$

The first expression gives $t' = At\big(1 - \frac{v}{c}\big)$. Substituting into the second:

$$ct = At'(c + v) = A\left(At\left(1 - \frac{v}{c}\right)\right)(c + v) = A^2t\left(c - \cancel{v} + \cancel{v} - \frac{v^2}{c}\right) = A^2t\left(c - \frac{v^2}{c}\right)$$

Solving for A:

$$\cancel{c}t = A^2\cancel{t}\left(c - \frac{v^2}{c}\right) \Rightarrow c^2 = A^2(c^2 - v^2) \Rightarrow A = \sqrt{\frac{c^2}{c^2 - v^2}} = \sqrt{\frac{1}{1 - \frac{v^2}{c^2}}}$$

Which is exactly the value y. An equation for time can be obtained in a similar fashion. In total, one obtains two systems of four equations each, called **Lorentz transformations**, that act as the relativistic counterpart to the Galilean transformations:

$$\begin{cases} x' = \gamma(x - vt) \\ y' = y \\ z' = z \\ t' = \gamma\big(t - \frac{v}{c^2}\big)x \end{cases} \qquad\qquad \begin{cases} x = \gamma(x' + vt') \\ y = y' \\ z = z' \\ t = \gamma\big(t' + \frac{v}{c^2}\big)x' \end{cases}$$

When the velocity v is much smaller than c, γ is close to 1 and the ratio vx/c^2 is close to 0, hence "falling back" to the Galilean transformations.

The last step in deriving the Lorentz transformations is an expression for velocities. The definition of velocity in the Galilean laws of motion is the derivative of space over the derivative of time. Carrying it out:

$$u' = \frac{dx'}{dt'} = \frac{d(\gamma(x - vt))}{d(\gamma(t - vx/c^2))} = \frac{\cancel{\gamma} d(x - vt)}{\cancel{\gamma} d(t - vx/c^2)} = \frac{dx - vdt}{dt - (v/c^2)dx} = \frac{\frac{dx}{dt} - v}{1 - (v/c^2)\frac{dx}{dt}}$$

But dx/dt is just u, giving:

$$u' = \frac{u - v}{1 - \frac{v}{c^2}u} \qquad\qquad u = \frac{u' + v}{1 + \frac{v}{c^2}u'}$$

When v is noticeably smaller than c, the ratio v/c^2 is infinitesimal, hence giving the same results as the Galilean transformations $u' = u - v$ and $u = u' + v$. Also, when $v = c$:

$$u' = \frac{u - c}{1 - \frac{\cancel{c}}{\cancel{c^2}}u} = \frac{u - c}{1 - \frac{u}{c}} = c\frac{\cancel{u - c}}{\cancel{c - u}} = -c \qquad\qquad u = \frac{u' + c}{1 + \frac{\cancel{c}}{\cancel{c^2}}u'} = \frac{u' + c}{1 + \frac{u'}{c}} = c\frac{\cancel{u' + c}}{\cancel{c + u'}} = c$$

Hence preserving the constancy of the speed of light in all reference frames.

Having introduced the relativistic γ factor also requires one to properly adapt not just Newtonian laws of motion, but also the definition of other physical quantities that depend on velocity, such as kinetic energy and momentum.

The classical definition of momentum is of mass times velocity; momentum is an important quantity in classical physics because it is always conserved. However, it can be shown that blindly applying such definition to particles where the Lorentz transformations are applied does not work, meaning that different frames of reference give different value of the momentum. So either momentum conservation should be abandoned, or the definition of momentum ought to be revisited.

Out of the two, the second is much more plausible. First, the definition of momentum of a particle is, as stated, mass times velocity, or mass times space over time. However, different frames of reference are associated to different time intervals and different lengths: the frame of reference of choice should be the one centered in the particle, since it's the one associated to proper time.

Let $u = \Delta x/\Delta t$ be the velocity of a particle in a frame of reference external to said particle. Let $\Delta\tau$ be the time measured in the frame of reference of the particle to describe the interlude of the same events described by Δt. Define the momentum of a particle with respect to its own frame of reference to be $p = mu$. Given that $\Delta\tau = \Delta t/\gamma$:

$$p = mu = m\frac{\Delta x}{\Delta\tau} = m\frac{\Delta x}{\Delta t/\gamma} = \gamma m\frac{\Delta x}{\Delta t} = \gamma mu$$

This is the expression for the relativistic momentum of a particle when observed to have velocity u from a given frame of reference. It can be proven that, as long as Lorentz transformations are used to transition from a frame of reference to another, the law of conservation of momentum still holds.

It is however important to note that the γ factor that appears in the expression for the relativistic momentum is somewhat different from the gamma factor that appears in the Lorentz transformations. The latter influences the velocity of a frame of reference with respect to another, the former influences the velocity of the particle *itself* with respect to the frame of reference in which it is observed. For this reason, to avoid distinction, the formula is sometimes written as $p = \gamma_p mu$, where the pedix p stands for "particle".

Velocity also appears in the expression for Newtonian kinetic energy, $K = \frac{1}{2}mu^2$. Since $p = mu$, the expression can also be written as $K = p^2/2m$. Having extended momentum to the relativistic case and its law of conservation, it is necessary to do the same with energy.

Consider a particle travelling a distance Δx in a time Δt, as observed from a given frame of reference S. As stated earlier, the spacetime interval $c^2\Delta^2 t - \Delta^2 x$ is invariant with respect to the frames of reference. Let $\Delta\tau$ be the time interval describing the same event when measured from the frame of reference of the particle itself.

Multiplying the expression for the spacetime interval by $(m/\Delta\tau)^2$ gives:

$$(c^2\Delta^2 t - \Delta^2 x)\left(\frac{m}{\Delta\tau}\right)^2 = \frac{m^2}{\Delta^2\tau}c^2\Delta^2 t - \frac{m^2}{\Delta^2\tau}\Delta^2 x = m^2c^2\frac{\Delta^2 t}{\Delta^2\tau} - m^2\frac{\Delta^2 x}{\Delta^2\tau} = (mc)^2\left(\frac{\Delta t}{\Delta\tau}\right)^2 - p^2$$

Now the expression refers to the relativistic momentum of the same particle when observed from two different reference frames. This quantity is also invariant, since both m and $\Delta\tau$ are constants.

Substituting $\Delta t = \gamma_p\Delta\tau$ and multiplying by c^2 gives:

$$c^2\left((mc)^2\left(\frac{\Delta t}{\Delta\tau}\right)^2 - p^2\right) = c^2(mc)^2\left(\frac{\gamma_p\cancel{\Delta\tau}}{\cancel{\Delta\tau}}\right)^2 - c^2p^2 = c^2\left(\gamma_p mc\right)^2 - c^2p^2 = \left(\gamma_p mc^2\right)^2 - (pc)^2$$

Since this quantity is invariant with respect to any frame of reference, computing it with respect to one frame of reference gives the value for any frame of reference. The most comfortable frame of reference to choose is the one of the particle itself, where $\gamma_p = 1$ and $p = 0$:

$$\left(1 \cdot mc^2\right)^2 - (0 \cdot c)^2 = \left(mc^2\right)^2 - 0^2 = \left(mc^2\right)^2$$

Giving:

$$\left(\gamma_p mc^2\right)^2 - (pc)^2 = \left(mc^2\right)^2$$

The expression involves three terms: $\gamma_p mc^2$, pc and mc^2 (all three squared). The second term is the (non relativistic) momentum of a particle multiplied by the speed of light, whereas the third term is its mass multiplied by the speed of light. Both of these are constants, since mass cannot change and momentum p is the one associated to a specific frame of reference, the one of the particle itself.

The first term is much more interesting, however. Since γ_p has no dimension, it has the same unit of measurement as mc^2: mass times velocity squared is energy. It is also not constant, since γ_p depends on the relative velocity of the particle. Applying binomial expansion:

$$\gamma_p mc^2 = \frac{mc^2}{\sqrt{1 - \frac{u^2}{c^2}}} \approx \left(1 + \frac{1}{2}\frac{u^2}{c^2}\right)mc^2 = mc^2 + \frac{1}{2}mu^2$$

The second term is clearly the (non relativistic) kinetic energy of the particle, whereas the first is a form of energy that does not depend on any property of the particle (its velocity, its position, ecc...) except for its mass. That is, simply by "existing" (almost all particles have a mass) a body possesses a form of energy.

This means that $\gamma_p mc^2$ represents the total energy of the particle. The total energy is given by a relativistic kinetic energy and a **rest energy**:

$$E = mc^2 + \left(\gamma_p - 1\right)mc^2 = E_0 + K$$

Since the relativistic kinetic energy depends on γ_p, it means that it's impossible for a particle's energy to grow indefinitely.

When p or u are 0, meaning that the particle is at rest, one has $E = mc^2$, which is a known result: the **mass-energy equivalence**.

This expression has meaning only for particles that have a mass; for massless particles like photons, it is convenient to substitute the newly found expression in the second expression, to get:

$$E^2 - (pc)^2 = E_0^2 \Rightarrow E = \sqrt{(pc)^2 + E_0^2}$$

In the case of massless particles, it is sufficient to set $E_0 = 0$ (since they have no mass, and hence no rest energy) to get $E = pc$. This implies an interesting result: it's possible to assign momentum to particles that have no mass. Also, even if all having mass equal to 0, massless particles moving at different velocities can have different momenta.

4. Quantum mechanics

4.1. Introduction

Quantum mechanics was a new paradigm developed in stages to answer questions that classical physics was unable to answer. In particular, it is a framework that is necessary to model reality at very small scales (atoms and molecules).

The first staple point of quantum mechanics is the idea that energy is not a *continuous* quantity, but is instead a *discrete* quantity, that is, an integer multiple of a fixed elementary value.

The fundamental physical constant that regulates the size of this fundamental energy bit is called the **Planck constant**, denoted as h:

$$h = 6.626 \times 10^{-34} J \cdot s$$

Since this value is very small, on large scales energy appears continuous, because the "steps" between different energy values are infinitesimal and become relevant only on small scales.

The second staple point of quantum mechanics is that the difference between particles and waves becomes blurred. In this sense, it is possible for matter particles to exhibit wave-like properties and it is possible for waves to exhibit particle-like properties.

4.1.1. Black body radiation

The first problem that quantum mechanics aided in solving is the description of the emitted radiation of a **black body**. A black body is an idealized physical body that is capable of absorbing any electromagnetic radiation, regardless of its frequency or angle of incidence, and that therefore emits back energy only and exclusively because of this absorbtion.

The spectrum of all the frequencies of electromagnetic radiations emitted from a black body, also referred to as **emission spectrum**, is given by:

$$I(f) = \frac{d^4 E}{d\theta dA \cos(\theta) dt df} = \frac{d^4 \Phi}{d\theta dA \cos(\theta) df}$$

Where E is the energy, A is the surface area, θ is the angle of emission, t is the timem f is the frequency and Φ is the flux. Since A and θ are effectively chosen by the experimenter, and are therefore always known, the only variable at play is the frequency (or the wavelength, which is just its reciprocal).

Experimentally, the emission spectrum forms a curved shape with a peak at a certain wavelenght $\lambda_{\max}$. It is possible to explicitly define the relationship between these two quantities as follows:

$$I(f, T) = \frac{2\pi h f^3}{c^2} \frac{1}{\exp\left(\frac{hf}{k_B T}\right) - 1}$$

Where T is the temperature, h is the Planck constant and k_B is the **Boltzmann constant**:

$$k_B = 1.38 \times 10^{-23} \frac{J}{\mathrm{K}}$$

The total energy emitted by the black body is given by:

$$E_{\text{tot}} = \sigma T^4$$

Where σ is the **Stefan-Boltzmann constant**:

$$\sigma 5.67 \times 10^{-8} \frac{W}{m^2 \mathrm{K}^4}$$

The formula suggests an intuitive result, mainly that the amount of energy emitted increases as the temperature increases, as vice versa.

It can be shown that the value of $\lambda_{\max}$ is inversely proportional to the temperature.

It is to be expected that the aforementioned equation for $I(f,T)$ were to be derived from Maxwell equations. Interestingly, there is no way to do so. To derive the expression for $I(f,T)$ it is necessary to assume that each particle acts as an harmonic oscillator that emits energy is chunks, not as a continuous stream.

Given a certain frequency, a single chunk of energy δE is given by:

$$\delta E = hf$$

4.1.2. Photoelectric effect

The second phenomena that quantum mechanics aided in explaining was the **photoelectric effect**. This phenomena is the expulsion of electrons, called *photoelectrons*, from a piece of metal when hit by light. This happens because the energy given to the plate by the light is sufficient to break the bond that links electrons to their nucleus, and are thus ejected.

It is possible to experimentally determine the number of electrons that are ejected from the plate and their energy with respect to the light frequency and intensity, and then derive from Maxwell equations the corresponding equations and see if the results match.

In the framework of classical mechanics, the energy transfer from light to electrons is no different than heating an object. This process requires time, so it is expected that when the plate is illuminated there should be some delay before the plate starts emitting electrons. Whan happens instead is that the electrons are emitted immediately, as soon as the plate is hit.

It should also be reasonable that an higher light intensity would correspond to an higher energy of the photoelectrons. Instead, the intensity of the light has no influence on the energy of the photoelectrons, which is instead proportional to the frequency of the light.

The third puzzling observation is that the photoelectric effect only happens when the incoming light has a frequency equal or above a certain threshold, specific for each metal. In classical mechanics this has no explanation, since the photoelectric effect should happen, albeit with different degree of intensity, when employing light of any frequency.

The quantum explanation is instead to assume that light is composed of elementary massless particles called **photons**, whose energy is given by the black body harmonic oscillator model. If this is the case, all three issues are solved, because:

1. If each electron is hit by photons one by one, there's no need to wait for the body to absorb energy, since energy absorbtion is "one-shot". Therefore, the electron explusion is instantaneous;
2. Since $E = hf$, energy is indeed dependent on the frequency;
3. If the absorbed energy is insufficient, the electron is immediately recaptured by the charge of the nucleus. Being the energy dependent on the frequency, this explains the existence of a frequency threshold.

4.1.3. Gasses and radiation

When a gas is traversed by light, it is expected that the overall resulting frequency of the light is lowered, but preserved. In a similar fashion, when light is induced to emit light (by a sparkle), it is

expected that every frequency is emitted. What is observed instead is that, for each gas, only specific frequencies are emitted/preserved.

The nuclear model of the atom was tested in an experiment by Rutherford. Of course, it is not possible to study atoms simply with a microscope, because the scale is too small even for the greatest magnifier. An alternative approach is to employ electric charges, since both electrons and protons are charged.

An extremely thin plate of gold is used as a probe. Gold is used because it is both very dens and very soft, and is therefore possible to craft extremely thin plates. A radioactive source emits alpha particles (which are just helium nuclei). A detector entirely surrounds the plate, so that it is possible to observe where (and if) alpha particles are deflected. What happens is that some particles, albeit in small number, are indeed deflected, sometimes with great angle, whereas in a non-nuclei model of the atom all particles would have stroke through.

The proper model of the atom was worked out by Rutherford, who imagined the atom as an incredibly dense nucleus of positive charges with electrons orbiting around it, so that the electromagnetic force between the two acts as a centripetal force. The deflection of the alpha particles are then caused by the electromagnetic repulsion of the nuclei when alpha particles get too close to them.

The problem with the model is that the electron orbiting around the nucleus, in order to be able to maintain its orbit, would emit radiation, which in turn means it would gradually loose angular momentum until it would spiral onto the nucleus. Even assuming this to be true, the time for this to happen would be too narrow for matter to exist.

A solution was to assume that the electron does indeed orbit around the nucleus, but can only find itself at very specific distances from it. For this to be possible, it is necessary to assume that the angular momentum of the electron is quantized, according to the formula:

$$|p| = mvr = n\frac{h}{2\pi} \text{ with } n \in \mathbb{N}$$

In particular, the energy of the electron can be found by employing Coulomb's Law and the expression for the electric potential:

$$E = K + U = -\frac{1}{4\pi\varepsilon_0}\frac{e^2}{2r}$$

Where e is the electric charge of the electron and r is its distance from the nucleus.

Being the energy quantized, the only non constant member of the equation, r, must also be quantized. This means that r cannot be lower than a certain threshold, which is determined as $0.53 \times 10^{-10}m$.

Each possible value of energy that an electron can possess is called an **energy level**. Each time an electron exchanges energy with the environment, it hops from one energy level to another. In particular, by releasing energy it goes down one level, by absorbing energy it goes up one level.

Let E_0 be the lowest possible energy level (the one associated to ($r = 0.53 \times 10^{-10}m$). It is possible to relate the energy of a generic level n with respect to E_0:

$$E_n = \frac{ke^2}{2n^2E_0}$$

Since $E = hf$, this explains why gasses can absorb/release only certain frequencies, because they are the ones that match the (fixed) energy amounts needed for electrons to move.