

Physics II.

György Hárs
Gábor Dobos

2013.04.30

Contents

1	Electrostatic phenomena - György Hárs	3
1.1	Fundamental experimental phenomena	3
1.2	The electric field	4
1.3	The flux	5
1.4	Gauss's law	6
1.5	Point charges and the Coulomb's law	7
1.6	Conservative force field	8
1.7	Voltage and potential	9
1.8	Gradient	10
1.9	Spherical structures	11
1.9.1	Metal sphere	11
1.9.2	Sphere with uniform space charge density	13
1.10	Cylindrical structures	15
1.10.1	Infinite metal cylinder	15
1.10.2	Infinite cylinder with uniform space charge density	17
1.11	Infinite parallel plate with uniform surface charge density	20
1.12	Capacitors	21
1.12.1	Cylindrical capacitor	22
1.12.2	Spherical capacitor	24
1.13	Principle of superposition	25
2	Dielectric materials - György Hárs	27
2.1	The electric dipole	27
2.2	Polarization	28
2.3	Dielectric displacement	30
2.4	Electric permittivity (dielectric constant)	32
2.5	Gauss's law and the dielectric material	32
2.6	Inhomogeneous dielectric materials	33
2.7	Demonstration examples	35
2.7.1	35
2.7.2	39

2.8	Energy relations	39
2.8.1	Energy stored in the capacitor	39
2.8.2	Principle of the virtual work	42
3	Stationary electric current (direct current) - György Hárs	44
3.1	Definition of Ampere	44
3.2	Current density (\mathbf{j})	45
3.3	Ohm's law	46
3.4	Joule's law	47
3.5	Microphysical interpretation	47
4	Magnetic phenomena in space - György Hárs	49
4.1	The vector of magnetic induction (\mathbf{B})	49
4.2	The Lorentz force	50
4.2.1	Cyclotron frequency	51
4.2.2	The Hall effect	53
4.3	Magnetic dipole	54
4.4	Earth as a magnetic dipole	57
4.5	Biot-Savart law	58
4.5.1	Magnetic field of the straight current	59
4.5.2	Central magnetic field of the polygon and of the circle	60
4.6	Ampere's law	61
4.6.1	Thick rod with uniform current density	63
4.6.2	Solenoid	64
4.6.3	Toroidal coil	65
4.7	Magnetic flux	66
5	Magnetic field and the materials - György Hárs	67
5.1	Three basic types of magnetic behavior	67
5.2	Solenoid coil with iron core	69
5.3	Ampere's law and the magnetic material	71
5.4	Inhomogeneous magnetic material	72
5.5	Demonstration example	75
5.6	Solenoid with iron core	77
6	Time dependent electromagnetic field - György Hárs	80
6.1	Motion related electromagnetic induction	80
6.1.1	Plane generator (DC voltage)	80
6.1.2	Rotating frame generator (AC voltage)	83
6.1.3	Eddy currents	84
6.2	Electromagnetic induction at rest	86

6.2.1	The mutual and the self induction	86
6.2.2	Induced voltage of a current loop	87
6.2.3	The transformer	88
6.2.4	Energy stored in the coil	93
6.3	The Maxwell equations	94
7	Electromagnetic oscillations and waves - Gábor Dobos	96
7.1	Electrical oscillators	96
7.2	Electromagnetic waves in perfect vacuum	99
7.3	Electromagnetic waves in non-conductive media	101
7.4	Direction of the E and B fields	102
7.5	Poynting Vector	102
7.6	Light-pressure	103
7.7	Skin depth	106
7.8	Reflection and refraction	108
8	Geometrical Optics - Gábor Dobos	111
8.1	Total internal reflection	111
8.2	Spherical Mirror	112
8.3	Thin spherical lenses	117
8.4	Projection by spherical lenses and mirrors	120
8.5	Aberrations	123
9	Wave optics - Gábor Dobos	127
9.1	Young's double slit experiment	127
9.2	Coherence	130
9.3	Multiple slit diffraction	132
9.4	Fraunhofer diffraction	137
9.5	Thin layer interference	142
10	Einstein's Special Theory of Relativity – Gábor Dobos	148
10.1	The Aether Hypothesis and The Michelson-Morley Experiment	148
10.2	Einstein's Special Theory of Relativity	152
10.3	Lorentz contraction and time dilatation	155
10.4	Velocity addition	159
10.5	Connection between relativistic and classical physics	160

Introduction

Present work is the summary of the lectures held by the author at Budapest University of Technology and Economics. Long verbal explanations are not involved in the text, only some hints which make the reader to recall the lecture. Refer here the book: Alonso/Finn Fundamental University Physics, Volume II where more details can be found.

Physical quantities are the product of a measuring number and the physical unit. In contrast to mathematics, the accuracy or in other words the precision is always a secondary parameter of each physical quantity. Accuracy is determined by the number of valuable digits of the measuring number. Because of this 1500 V and 1.5 kV are not equivalent in terms of accuracy. They have 1 V and 100 V absolute errors respectively. The often used term relative error is the ratio of the absolute error over the nominal value. The smaller is the relative error the higher the accuracy of the measurement. When making operations with physical quantities, remember that the result may not be more accurate than the worst of the factors involved. For instance, when dividing 3.2165 V with 2.1 A to find the resistance of some conductor, the result 1.5316667 ohm is physically incorrect. Correctly it may contain only two valuable digits, just like the current data, so the correct result is 1.5 ohm.

The physical quantities are classified as fundamental quantities and derived quantities. The fundamental quantities and their units are defined by standard or in other words etalon. The etalons are stored in relevant institute in Paris. The fundamental quantities are the length, the time and the mass. The corresponding units are meter (m), second (s) and kilogram (kg) respectively. These three fundamental quantities are sufficient to build up the mechanics. The derived quantities are all other quantities which are the result of some kind of mathematical operations. To describe electric phenomena the fourth fundamental quantity has been introduced. This is ampere (A) the unit of electric current. This will be used extensively in Physics 2, when dealing with electricity.

Chapter 1

Electrostatic phenomena - György Hárs

1.1 Fundamental experimental phenomena

Electrostatics deals with the phenomena of electric charges at rest. Electric charges can be generated by rubbing different insulating materials with cloth or fur. The device called electroscope is used to detect and roughly measure the electric charge. By rubbing a glass rod and connecting it to the electroscope the device will indicate that charge has been transferred to it. By doing so second time the electroscope will indicate even more charges. Accordingly the same polarity charges are added together and are accumulating on the electroscope. Now replace the glass rod with a plastic rod. If the plastic rod is rubbed and connected to the already charged electroscope, the excursion of the electroscope will decrease. This proves that there are two opposite polarity charges in the nature, therefore they neutralize each other. The generated electricity by glass and plastic are considered positive and negative, respectively. The unit of the charge is called “coulomb” which is not fundamental quantity in System International (SI) so “ampere second” (As) is used mostly.

Now take a little (roughly 5 mm in diameter) ball of a very light material and hang it on a thread. This test device is able to detect forces by being deflected from the vertical. Charge up the ball to positive and approach it with a charged rod. If the rod is positive or negative the force is repulsive or attractive, respectively. This experiment demonstrates that the opposite charges attract, the same polarity charges repel each other.

Now use neutral test device in the next experiment. Put the ball to close proximity of the rod with charge on it. The originally neutral ball will be attracted. By approaching the rod with the ball even more the ball will suddenly be repelled once mechanically connected. The explanation of this experiment is based on the phenomenon of electrostatic induction (or some say electrostatic influence). By the effect of the external charge the

neutral ball became a dipole. For the sake of simplicity assume positive charge on the rod. The surface closer to the rod is turned to negative, while the opposite side became positive. The attractive force of the opposite charges is higher (due to the smaller distance) than the repelling force of the other side. So altogether the ball will experience a net attractive force. When the rod connected to the ball it became positively charged and was immediately repelled.

1.2 The electric field

In the proximity of the charged objects forces are exerted to other charges. The charge under investigation is called the “source charge”. To map the forces around the source charge a hypothetic positive point-like charge is used which is called the “test charge” denoted with q . By means of the test charge the force versus position function can be recorded. In terms of mathematics this is a vector-vector function or in other words force field $\mathbf{F}(\mathbf{r})$. Experience shows that the intensity of force is linearly proportional with the test charge. By dividing the force field with the amount of the test charge one recovers a normalized parameter. This parameter is the electric field $\mathbf{E}(\mathbf{r})$ which is characteristic to electrification state of the space generated solely by the source charge. The unit of electric field is N/As or much rather V/m . In Cartesian coordinates the vector field consists of three pieces of three variable functions.

$$\frac{\mathbf{F}(\mathbf{r})}{q} = \mathbf{E}(\mathbf{r}) = E_x(x, y, z)\mathbf{i} + E_y(x, y, z)\mathbf{j} + E_z(x, y, z)\mathbf{k} \quad (1.1)$$

One variable scalar functions $y = f(x)$ are easy to display in Cartesian system as curve. In case of two or three independent variables a scalar field is generated. This can be displayed like level curves or level surfaces. To display the vector field requires the concept of force line. Force lines are hypothetic lines with the following criteria:

- Tangent of the force line is the direction of the force vector
- Density of the force lines is proportional with the absolute value (intensity) of the vector.

The positive test charge is repelled by the positive source charge therefore the electric field $\mathbf{E}(\mathbf{r})$ lines are virtually coming out from the positive source charge. One might say that the positive charge is the source of the electric field lines. (The outcome would be the same by assuming negative test charge, this time the force would be opposite but after division with the negative test charge the direction of the electric field would revert.) The negative source charge is the drain of the electric field lines due to symmetry reasons. So the electric field lines start on the positive charge and end on the negative charge. When both positive and negative charges are present in the space the electric field lines

leaving the positive charges are drained fully or partially by the negative charges. The electric field lines of the uncompensated positive or negative charge will end or start in the infinity, respectively.

In case when more source charges are present in the empty space the principle of superposition is valid. Accordingly the electric field vectors are added together as usual vector addition in physics.

1.3 The flux

To understand the concept of flux we start with a simple example and proceed to the general arrangement.

Assume we have a tube with stationary flow of water in which the velocity versus position vector field $\mathbf{v}(\mathbf{r})$ is homogeneous, in other words the velocity vector is constant everywhere. Now take a plane-like frame made of a very thin wire with the area vector \mathbf{A} . The area vector by definition is normal to the surface and the absolute value of the vector is the area of the surface. Let us submerge the frame into the flowing water. The task is to find a formula for the amount of water going through the frame.

If the area vector is parallel with the velocity (this means that the velocity vector is normal to the surface) the flow rate (Φ) through the frame (m^3/s) is simply the product of the area and the velocity. If the angle between the area vector and the velocity is not zero but some other φ angle, the area vector should be projected to the direction of the velocity. The projection can be carried out by multiplying with the cosine of the angle. So ultimately it can be stated that the flow rate is the dot product of the velocity vector and the area vector.

$$\Phi = \mathbf{v} \cdot \mathbf{A}. \quad (1.2)$$

Remember that the above simple formula is valid in case of homogeneous vector field and plane-like frame alone. The question is how the above argument can be implemented to the general case where the vector field is not homogeneous and the frame has a curvy shape. The solution requires subdividing the area to very small mosaics which represent the surface like tiles on a curvy wall. If the mosaics are sufficiently small (math says they are infinitesimal) then the vector field can be considered homogeneous within the mosaic, and the mosaic itself can be considered plain. So ultimately the above simple dot product can be readily used for the little mosaic. At each mosaic one has to choose a representing value of the velocity vector since the velocity vector changes from place to place. The surface vector also changes from point to point, since the surface is not plane-like any more. Finally the contribution of each mosaic has to be summarized. If the process of subdivision goes to the infinity than the summarized value tends to a limit which is called flux, or in terms of mathematics it is called the scalar value surface

integral (denoted as below).

$$\Phi = \int_S \mathbf{v}(\mathbf{r}) \cdot d\mathbf{A} \quad (1.3)$$

Here S indicates an open surface on which the integration should be carried out. The open surface has a rim and has two sides just like a sheet of paper. In contrast to it, the closed surface does not have any rim and divides the 3D space to internal and external domains just like a ball. In case of open surface the circulation of the rim determines the direction of the area vector like a right hand screw turning. Since at closed surface there is no rim a convention states that the area vector is directed outside direction.

Let us find out how much the above integral would be if a closed surface would be submerged into the flow of water, of course with a penetrable surface. In physical context the fact is clear that on one side of the surface the water flows in and on the other side it flows out. After some consideration one can readily conclude that the overall flux on a closed surface is zero. This statement is true as long as the closed surface does not contain source or drain of the water. If the closed surface contains source then the velocity vectors all point away from the surface, thus the flux will be a positive value equal to the intensity of the source. Plausibly negative result comes out when the drain is contained by the surface. This time the negative value is the intensity of the drain enclosed. The integral to a closed surface is denoted as follows:

$$\Phi = \oint_S \mathbf{v}(\mathbf{r}) \cdot d\mathbf{A} \quad (1.4)$$

1.4 Gauss's law

In section 1.2. the fact has been stated that positive and negative charge are the source and the drain of the electric field lines, respectively. Combining this with the features flux on a closed surface the conclusion is clear: The flux of the electric field to a close surface is zero as long as the surface does not contain charge. When it does contain charge the flux will be proportional with the amount of charge enclosed. If the charge is positive or negative the flux will be the same sign value. In terms of formula this is the Gauss's law:

$$\oint_S \mathbf{E}(\mathbf{r}) \cdot d\mathbf{A} = \frac{Q}{\epsilon_0} \quad (1.5)$$

On the right hand side Q denotes the total charge contained by the surface, vacuum permittivity ϵ_0 is a universal constant in nature. ($\epsilon_0 = 8.86 \cdot 10^{-12} \text{As/Vm}$)

We want to use Gauss's law for solving problems in which the charge arrangement is given and the distribution of the electric field is to be found. This law is an integral type law. In general case information is lost by integration. The only case when information is preserved is when the function to be integrated is constant. Therefore there will be three distinct classes of charge arrangements when the Gauss's law can be effectively used. These are as follows:

- Spherically symmetric
- Cylindrically symmetric, infinite long
- Plane parallel, infinite large

In all other cases the Gauss's law is also true in terms of integral, but the local electric field is impossible to determine.

To use the law actually, one needs a closed surface with the same symmetry as that of the charge arrangement. On this surface the angle of electric field vector is necessarily normal and its intensity is constant. This way the vector integral of flux is majorly simplified to the product of the area and the electric field intensity.

1.5 Point charges and the Coulomb's law

Let us use the Gauss's law for the case of point charge. Point charge is a model with zero extension and finite (non infinitesimal) charge. Accordingly the charge density and the electrostatic energy are infinite. Even though, this is a useful model for many charge arrangements which are much larger than the distinct charges themselves.

The electric field is perfectly spherical around the point charge. So surface to be used is obviously sphere. The surface of the sphere is $4r^2\pi$. Accordingly Gauss's law can be written as follows:

$$4r^2\pi \cdot E = \frac{Q}{\varepsilon_0} \quad (1.6)$$

The electric field can readily be expressed:

$$E = \frac{Q}{4\pi\varepsilon_0} \cdot \frac{1}{r^2} \quad (1.7)$$

The above formula is the electric field of the point charge which will be used extensively later in this chapter.

The exerted force to a q charge can be written:

$$F = \frac{1}{4\pi\varepsilon_0} \cdot \frac{Qq}{r^2} \quad (1.8)$$

This is the Coulomb's law which describes the force between point charges. For practical reason it is worth remembering that the value of the constant in Coulomb's law is the following:

$$k = \frac{1}{4\pi\epsilon_0} = 9 \cdot 10^9 \frac{Vm}{As} \quad (1.9)$$

1.6 Conservative force field

Force field is a vector-vector function in which the force vector \mathbf{F} depends on the position vector \mathbf{r} . In terms of mathematics the force field $\mathbf{F}(\mathbf{r})$ is described as follows:

$$\mathbf{F}(\mathbf{r}) = X(x, y, z)\mathbf{i} + Y(x, y, z)\mathbf{j} + Z(x, y, z)\mathbf{k} \quad (1.10)$$

where \mathbf{i} , \mathbf{j} , \mathbf{k} are the unit vectors of the coordinate system.

Take a test charge and move it slowly in the $\mathbf{F}(\mathbf{r})$ force field from position \mathbf{A} to position \mathbf{B} on two alternative paths.

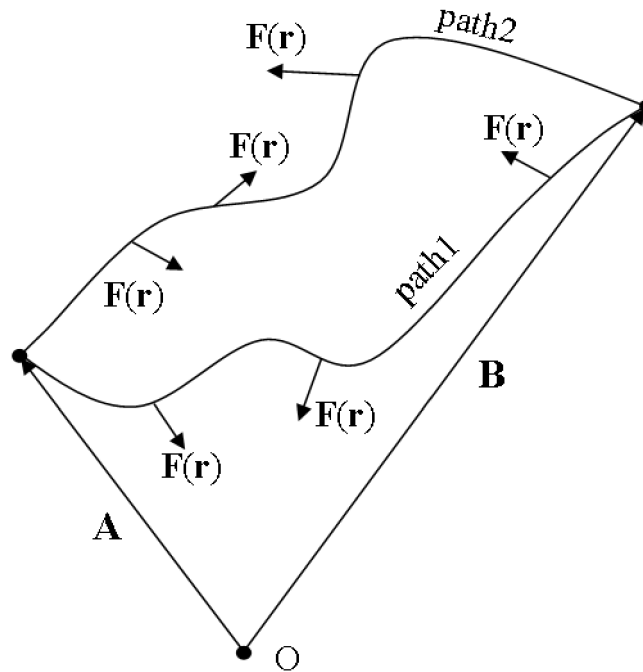


Figure 1.1: Integration on two paths

Let us calculate the amount of work done on each path. The force exerted to the test charge by my hand is just opposite of the force field $-\mathbf{F}(\mathbf{r})$. If it was not the case, the charge would accelerate. The moving is thought to happen quasi-statically without acceleration.

Let us calculate my work for the two alternate paths:

$$W_1 = \left(\int_{\mathbf{A}}^{\mathbf{B}} (-\mathbf{F}) d\mathbf{r} \right)_{path1} \quad W_2 = \left(\int_{\mathbf{A}}^{\mathbf{B}} (-\mathbf{F}) d\mathbf{r} \right)_{path2} \quad (1.11)$$

In general case W_1 and W_2 are not equal. However, in some special cases they may be equal for any two paths. Imagine that our force field is such, that W_1 and W_2 are equal. In this case a closed loop path can be made which starts with path 1 and returns to the starting point on path 2. Since the opposite direction passage turns W_2 to its negative, ultimately the closed loop path will result in zero. That special force field where the integral is zero for any closed loop is considered CONSERVATIVE force field. In formula:

$$\oint \mathbf{F}(\mathbf{r}) d\mathbf{r} = 0 \quad (1.12)$$

Using the concept of electric field with the formula $\mathbf{F} = q\mathbf{E}$ above equation is transformed:

$$\oint \mathbf{E}(\mathbf{r}) d\mathbf{r} = 0 \quad (1.13)$$

According to the experience the electric field obeys the law of conservative field. The integral on any closed loop results zero. That also means that curve integral between any two points is independent of the path and solely depends on the starting and final point.

1.7 Voltage and potential

The work done against the force of the electric field is as follows:

$$W_{\mathbf{A}}^{\mathbf{B}} = \int_{\mathbf{A}}^{\mathbf{B}} (-\mathbf{F}) d\mathbf{r} = q \int_{\mathbf{A}}^{\mathbf{B}} (-\mathbf{E}) d\mathbf{r} \quad (1.14)$$

Let us rearrange and divide with q .

$$U_{\mathbf{A}}^{\mathbf{B}} = \frac{W_{\mathbf{A}}^{\mathbf{B}}}{q} \quad U_{\mathbf{A}}^{\mathbf{B}} = - \int_{\mathbf{A}}^{\mathbf{B}} \mathbf{E}(\mathbf{r}) d\mathbf{r} \quad (1.15)$$

The voltage of point \mathbf{B} relative to \mathbf{A} is given by the formula above. The fact is clear that the voltage is dependent on two points. If the starting point is considered as a reference

point for all the integrals, that specific voltage will be dependent on the final point only. This one parameter voltage is called the potential.

$$U_{\mathbf{B}} = - \int_{\mathbf{ref}}^{\mathbf{B}} \mathbf{E}(\mathbf{r}) d\mathbf{r} \quad (1.16)$$

Voltages can be expressed as the difference of potentials proven below:

$$U_{\mathbf{A}}^{\mathbf{B}} = \left(- \int_{\mathbf{A}}^{\mathbf{ref}} \mathbf{E}(\mathbf{r}) d\mathbf{r} \right) + \left(- \int_{\mathbf{ref}}^{\mathbf{B}} \mathbf{E}(\mathbf{r}) d\mathbf{r} \right) = \left(- \int_{\mathbf{ref}}^{\mathbf{B}} \mathbf{E}(\mathbf{r}) d\mathbf{r} \right) - \left(- \int_{\mathbf{ref}}^{\mathbf{A}} \mathbf{E}(\mathbf{r}) d\mathbf{r} \right) = U_{\mathbf{B}} - U_{\mathbf{A}} \quad (1.17)$$

The potential of any point can also be written as follows:

$$U(\mathbf{r}) = - \int_{\mathbf{ref}}^{\mathbf{r}} \mathbf{E}(\mathbf{r}') d\mathbf{r}' \quad (1.18)$$

The concept of voltage exists always and its value is definite. The value of potential is indefinite because it depends on the reference point too. However this is possible to define a definite potential. The reference point of the integral should be placed to the infinity. This can be done in case of physically real objects when the corresponding improper integral is convergent. The physically real object is by definition such an object which virtually shrinks to a point if one departs infinite far away. The spherical charge arrangement is the only physically real object among those mentioned above. An infinite long cylinder or an infinite plan parallel plate are not physically real, since viewed from infinite it still looks infinite.

1.8 Gradient

The gradient is an operation in vector calculus which generates the electric field vector from the potential scalar field. In general case the formula is as follows:

$$\text{grad}U(\mathbf{r}) = \frac{\partial U(\mathbf{r})}{\partial x} \mathbf{i} + \frac{\partial U(\mathbf{r})}{\partial y} \mathbf{j} + \frac{\partial U(\mathbf{r})}{\partial z} \mathbf{k} = -\mathbf{E}(\mathbf{r}) \quad (1.19)$$

Therefore:

$$\mathbf{E}(\mathbf{r}) = -\text{grad}U(\mathbf{r}) \quad (1.20)$$

In the special case of spherical cylindrical and plane parallel structures the gradient operation is merely a derivation according to the position variable.

$$E(r) = - \frac{dU(r)}{dr} \quad (1.21)$$

1.9 Spherical structures

1.9.1 Metal sphere

Metal sphere with radius $R = 0.1m$ contains $Q = 10^{-8}As$ charge. Find the function of the electric field and the potential as the function of distance from the center and sketch the result. Calculate the values of the electric field and the potential on the surface of the metal sphere. Determine the capacitance of the metal sphere.

Metal contains free electrons therefore electric field may not exist inside the bulk of the metal. If there was electric field in the metal the free electrons would move to compensate it to zero very fast. Since there is no electric field inside the metal the total volume of the metal is equipotential. The vector of the electric field is always normal (perpendicular) to the metal (equipotential) surface. The proof of this as follows: If there was an angle different of ninety degrees then this electric field vector could be decomposed to normal and tangential components. The tangential component would readily move the electrons until this component gets compensated. In stationary case all the excess charge resides on the surface of the metal. Therefore a hollow metal is equivalent with a bulky metal in terms of electrostatics. The surface charge density and the surface electric field are proportional to the reciprocal of the curvature radius.

$$\oint_S \mathbf{E}(\mathbf{r}) \cdot d\mathbf{A} = \frac{Q}{\varepsilon_0} \quad (1.22)$$

Gauss's law is used to solve the problem. We pick a virtual point-like balloon and inflate it from zero to the infinity radius. Inside the metal sphere there is no contained charge in the balloon.

$$4r^2\pi \cdot E = 0 \quad (1.23)$$

$$E = 0 \quad (1.24)$$

So the electric field inside the metal sphere is zero.

Out of the metal sphere however the contained charge is the amount given in this problem.

$$4r^2\pi \cdot E = \frac{Q}{\varepsilon_0} \quad (1.25)$$

The electric field can be expressed:

$$E(r) = \frac{Q}{4\pi\varepsilon_0} \cdot \frac{1}{r^2} \quad (1.26)$$

On the surface of the metal sphere the electric field comes out if $r = R$ is substituted to the above function

$$E(R) = \frac{Q}{4\pi\epsilon_0} \cdot \frac{1}{R^2} = 9 \cdot 10^9 \frac{10^{-8}}{0.1^2} = 9000 \frac{V}{m} \quad (1.27)$$

The potential function can be determined by integrating the electric field:

$$U(r) = - \int_{\infty}^r E(r') dr' = - \int_{\infty}^r \frac{Q}{4\pi\epsilon_0} \cdot \frac{1}{r'^2} dr' = \frac{Q}{4\pi\epsilon_0} \cdot \int_{\infty}^r \left(-\frac{1}{r'^2}\right) dr' = \quad (1.28)$$

$$= \frac{Q}{4\pi\epsilon_0} \left[\frac{1}{r'} \right]_{r'=\infty}^{r'=r} = \frac{Q}{4\pi\epsilon_0} \left(\frac{1}{r} - \frac{1}{\infty} \right) = \frac{Q}{4\pi\epsilon_0} \frac{1}{r} \quad (1.29)$$

So briefly the potential function out of the metal sphere is as follows:

$$U(r) = \frac{Q}{4\pi\epsilon_0} \frac{1}{r} \quad (1.30)$$

On the surface of the metal sphere the potential comes out if $r = R$ is substituted to the above function

$$U(R) = \frac{Q}{4\pi\epsilon_0} \frac{1}{R} = 9 \cdot 10^9 \frac{10^{-8}}{0.1} = 900V \quad (1.31)$$

Inside the metal sphere the potential is constant due to the zero electric field.

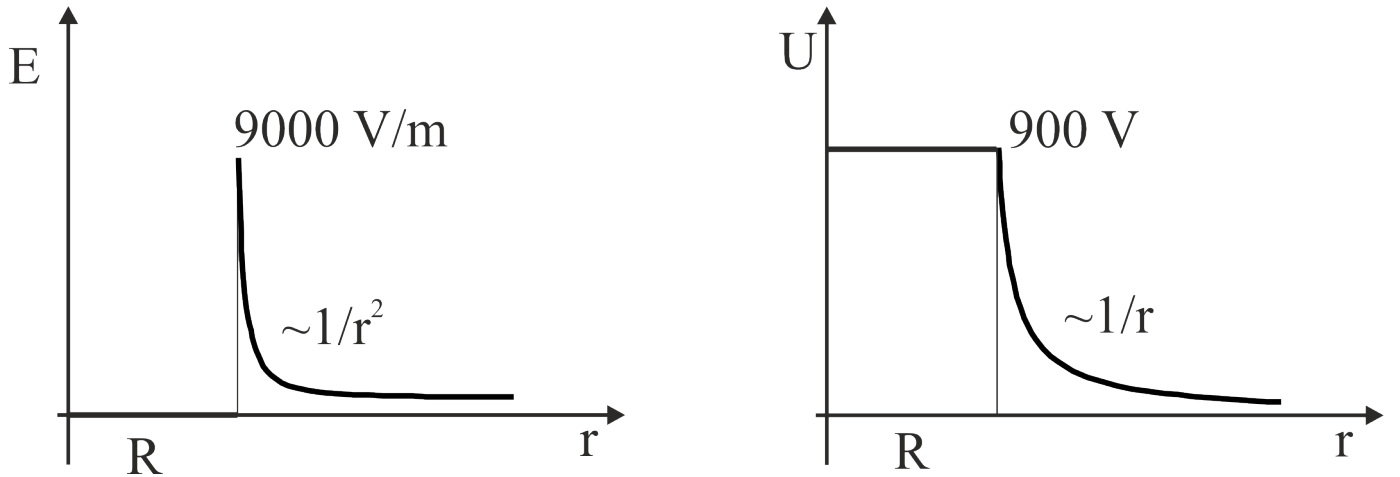


Figure 1.2: Metal sphere
 Electric field vs. radial position function Potential vs. radial position function

An interesting result can be concluded. Let us divide the formula of the potential and the electric field on the surface.

$$\frac{U(R)}{E(R)} = \frac{Q}{4\pi\epsilon_0} \frac{1}{R} \cdot \frac{4\pi\epsilon_0}{Q} R^2 = R \quad (1.32)$$

The electric field on the surface is the ratio of the potential and the radius.

$$E(R) = \frac{U(R)}{R} \quad (1.33)$$

The result is in perfect agreement with the numerical values.

This result is useful when high electric field is desired. This time ultra sharp needle is used and the needle is hooked up to high potential. By means of this device corona discharge can be generated in air.

Capacitance is a general term in physics which means a kind of storage capability. More precisely this is the ratio of some kind of extensive parameter over the corresponding intensive parameter. For instance the heat capacitance is the ratio of the heat energy over the temperature. Similarly the electric capacitance is the ratio of the charge over the generated potential. The unit of the capacitance is As/V which is called Farad (F) to commemorate the famous scientist Faraday. Farad as a unit is very large therefore pF or μF is used mostly. Capacitance denoted with C is a feature of all physically real conductive objects. In contrast to this the capacitor is a device used in the electronics with intentionally high capacitance.

$$U = \frac{Q}{4\pi\epsilon_0} \frac{1}{R} \quad (1.34)$$

$$C = \frac{Q}{U} = 4\pi\epsilon_0 R = \frac{1}{9 \cdot 10^9} \cdot 1Farad = 110pF \quad (1.35)$$

So the capacitance of the metal sphere is proportional to the radius. It is worth remembering that a big sphere of one meter radius has a capacitance of 110 pF . The capacitance of the human body is in the range of some tens of pF .

1.9.2 Sphere with uniform space charge density

Uniform space charge density ($\rho = 10^{-6} As/m^3$) is contained by a sphere with radius $R = 0.1$ meter. (The charge density is immobile. Imagine this in the way that wax is melted charged up and let it cool down. The charges are effectively trapped in the wax.) Find the function of the electric field and the potential as the function of distance from

the center and sketch the result. Calculate the value of the electric field on the surface of the sphere and the value of the potential on the surface and in the center.

$$\oint_S \mathbf{E}(\mathbf{r}) \cdot d\mathbf{A} = \frac{Q}{\varepsilon_0} \quad (1.36)$$

We pick a virtual point-like balloon and inflate it from zero to the infinity radius. Inside the charged sphere the Gauss's law is as follows:

$$4r^2\pi \cdot E = \frac{4r^3\pi}{3} \frac{\rho}{\varepsilon_0} \quad (1.37)$$

On the left hand side there is the flux on the right hand side there is the volume of the sphere multiplied with the charge density. Many terms cancel out.

$$E(r) = \frac{\rho}{3\varepsilon_0} r \quad (1.38)$$

The result is not surprising. By increasing the radius in the sphere the charge contained grows cubically the surface area increases with the second power so the ratio will be linear.

Outside the charged sphere the amount of the charge contained does not grow any more only the surface of the sphere continues to grow with the second power.

$$4r^2\pi \cdot E = \frac{4R^3\pi}{3} \frac{\rho}{\varepsilon_0} \quad (1.39)$$

$$E(r) = \frac{\rho}{3\varepsilon_0} \frac{R^3}{r^2} \quad (1.40)$$

The two above equations show that the function of the electric field is continuous, since on the surface of the charged sphere $r = R$ substitution produces the same result.

On the surface of the sphere the numerical value of the electric field can readily be calculated:

$$E(R) = \frac{\rho}{3\varepsilon_0} R = \frac{10^{-6}}{3 \cdot 8,86 \cdot 10^{-12}} \cdot 0.1 = 3762 \frac{V}{m} \quad (1.41)$$

The potential function can be determined by integrating the electric field. First the external region is integrated:

$$U_{out}(r) = - \int_{\infty}^r E(r') dr' = - \int_{\infty}^r \frac{\rho}{3\varepsilon_0} \frac{R^3}{r'^2} dr' = \quad (1.42)$$

$$= \frac{\rho R^3}{3\varepsilon_0} \cdot \int_{\infty}^r \left(-\frac{1}{r'^2}\right) dr' = \frac{\rho R^3}{3\varepsilon_0} \left[\frac{1}{r'} \right]_{r'=\infty}^{r'=r} = \frac{\rho R^3}{3\varepsilon_0} \left(\frac{1}{r} - \frac{1}{\infty} \right) = \frac{\rho R^3}{3\varepsilon_0} \frac{1}{r} \quad (1.43)$$

So briefly the potential function out of the charged sphere is as follows:

$$U_{out}(r) = \frac{\rho R^3}{3\varepsilon_0} \frac{1}{r} \quad (1.44)$$

The surface potential of the sphere is the above function with $r = R$ substitution:

$$U(R) = \frac{\rho R^2}{3\varepsilon_0} = \frac{10^{-6} \cdot 10^{-2}}{3 \cdot 8,86 \cdot 10^{-12}} = 376V \quad (1.45)$$

Remember that this value should be added to the integral calculated next.

Inside the charged sphere the integral is different:

$$U_{in}(r) = U(R) + U_R^r = U(R) + \left(- \int_R^r E(r') dr' \right) \quad (1.46)$$

For simplicity reason only the integral in the parenthesis is transformed first:

$$U_R^r = - \int_R^r \frac{\rho}{3\varepsilon_0} r' dr' = - \frac{\rho}{3\varepsilon_0} \int_R^r r' dr' = - \frac{\rho}{3\varepsilon_0} \left[\frac{r'^2}{2} \right]_{r'=R}^{r'=r} = - \frac{\rho}{3\varepsilon_0} \left(\frac{r^2}{2} - \frac{R^2}{2} \right) \quad (1.47)$$

Altogether:

$$U_{in}(r) = U(R) + U_R^r = \frac{\rho R^2}{3\varepsilon_0} + \left(- \frac{\rho}{3\varepsilon_0} \left(\frac{r^2}{2} - \frac{R^2}{2} \right) \right) = \frac{\rho}{3\varepsilon_0} \left(\frac{3R^2}{2} - \frac{r^2}{2} \right) = \frac{\rho}{6\varepsilon_0} (3R^2 - r^2) \quad (1.48)$$

The final result is:

$$U_{in}(r) = \frac{\rho}{6\varepsilon_0} (3R^2 - r^2) \quad (1.49)$$

The numerical value of the central potential is given by the above equation at $r = 0$ substitution.

$$U_{in}(0) = \frac{\rho}{6\varepsilon_0} (3R^2) = \frac{\rho R^2}{2\varepsilon_0} = \frac{10^{-6} \cdot 10^{-2}}{2 \cdot 8,86 \cdot 10^{-12}} = 564V \quad (1.50)$$

1.10 Cylindrical structures

1.10.1 Infinite metal cylinder

Infinite metal cylinder (tube) with radius $R = 0.1m$ contains $\omega = 10^{-8} As/m^2$ surface charge density. Find the function of the electric field and the potential as the function

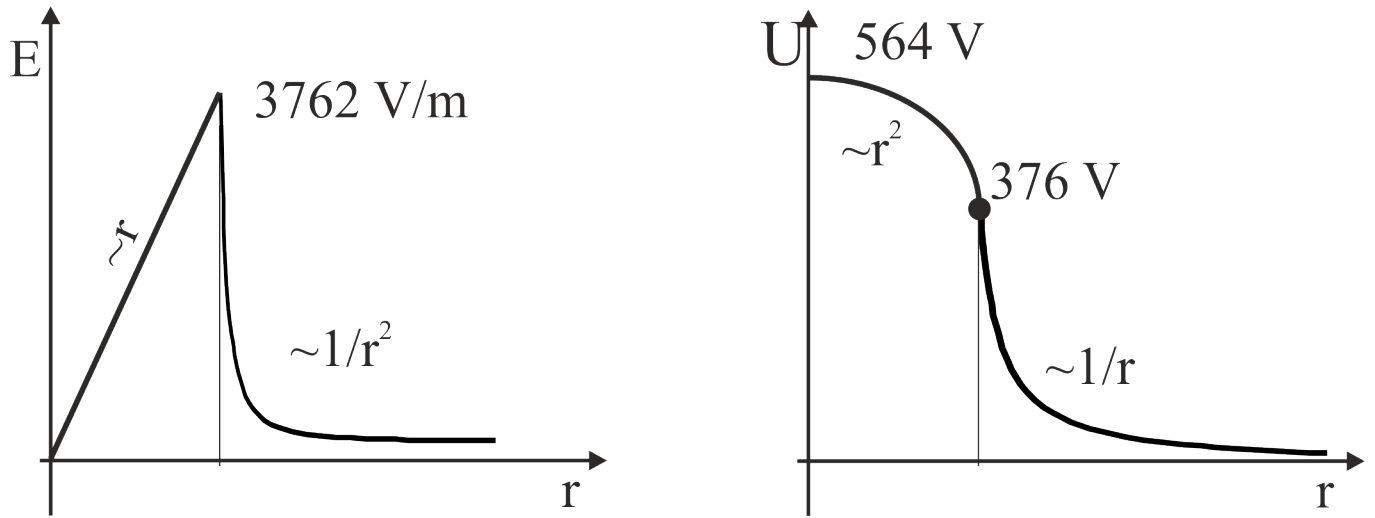


Figure 1.3: Sphere with uniform charge density
 Electric field vs. radial position function Potential vs. radial position function

of distance from the center and sketch the result. The reference point of the potential should be the center. Calculate the value of the electric field and of the potential on the surface of the metal cylinder.

$$\oint_S \mathbf{E}(\mathbf{r}) \cdot d\mathbf{A} = \frac{Q}{\varepsilon_0} \quad (1.51)$$

Gauss's law is used to solve the problem. We pick a virtual line-like tube and inflate it from zero to the infinity radius. Inside the metal cylinder there is no contained charge in the virtual tube.

$$2r\pi \cdot l \cdot E = 0 \quad (1.52)$$

$$E = 0 \quad (1.53)$$

So the electric field inside the metal cylinder is zero.

Out of the metal cylinder however the contained charge is as follows.

$$2r\pi \cdot l \cdot E = \frac{2R\pi \cdot l \cdot \omega}{\varepsilon_0} \quad (1.54)$$

The electric field can be expressed:

$$E(r) = \frac{\omega R}{\varepsilon_0} \cdot \frac{1}{r} \quad (1.55)$$

On the surface of the metal cylinder the electric field comes out if $r = R$ is substituted to the above function

$$E(R) = \frac{\omega R}{\varepsilon_0} \cdot \frac{1}{R} = \frac{\omega}{\varepsilon_0} = \frac{10^{-8}}{8.86 \cdot 10^{-12}} = 1129 \frac{V}{m} \quad (1.56)$$

The potential function can be determined by integrating the electric field:

$$U(r) = - \int_R^r E(r') dr' = - \int_R^r \frac{\omega R}{\varepsilon_0} \cdot \frac{1}{r'} dr' = - \frac{\omega R}{\varepsilon_0} \int_R^r \frac{1}{r'} dr' = - \frac{\omega R}{\varepsilon_0} [\ln r']_{r'=R}^{r'=r} = - \frac{\omega R}{\varepsilon_0} \ln \left(\frac{r}{R} \right) \quad (1.57)$$

So briefly the potential function out of the metal cylinder is as follows:

$$U(r) = - \frac{\omega R}{\varepsilon_0} \ln \left(\frac{r}{R} \right) \quad (1.58)$$

On the surface of the metal cylinder the potential comes out if $r = R$ is substituted to the above function

$$U(R) = - \frac{\omega R}{\varepsilon_0} \ln \left(\frac{R}{R} \right) = 0V \quad (1.59)$$

The result is obvious since inside the metal cylinder the potential is constant due to the zero electric field.

Note that the reference point of the potential could not be placed to the infinity because the infinite long cylinder is not physically real object. Therefore the improper integral is not convergent.

1.10.2 Infinite cylinder with uniform space charge density

Uniform space charge density ($\rho = 10^{-6} \text{ As/m}^3$) is contained by an infinite cylinder with radius $R = 0.1$ meter. (The charge density is immobile. Imagine this in the way that wax is melted charged up and let it cool down. The charges are effectively trapped in the wax.) Find the function of the electric field and the potential as the function of distance from the center and sketch the result. Calculate the value of the electric field and the potential on the surface of the cylinder. The reference point of the potential should be the central line.

$$\oint_S \mathbf{E}(\mathbf{r}) \cdot d\mathbf{A} = \frac{Q}{\varepsilon_0} \quad (1.60)$$

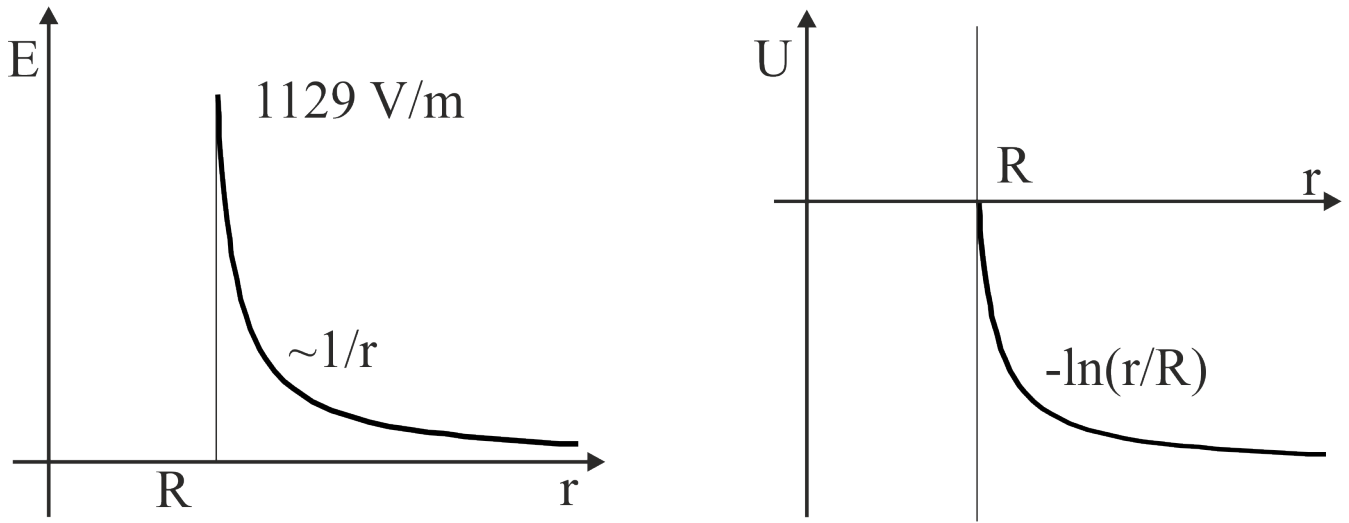


Figure 1.4: Metal cylinder

Electric field vs. radial position function

Potential vs. radial position function

We pick a virtual line-like tube and inflate the radius from zero to the infinity. Inside the charged sphere the Gauss's law is as follows:

$$2r\pi \cdot l \cdot E = r^2\pi \cdot l \frac{\rho}{\varepsilon_0} \quad (1.61)$$

On the left hand side there is the flux on the right hand side there is the volume of the cylinder multiplied with the charge density. Many terms cancel out.

$$E(r) = \frac{\rho}{2\varepsilon_0} r \quad (1.62)$$

The result is not surprising. By increasing the radius the charge contained grows with the second power, the surface area increases linearly so the ratio will be linear.

Outside the charged cylinder the amount of the charge contained does not grow any more only the surface continues to grow linearly.

$$2r\pi \cdot l \cdot E = R^2\pi \cdot l \frac{\rho}{\varepsilon_0} \quad (1.63)$$

$$E(r) = \frac{\rho}{2\varepsilon_0} \frac{R^2}{r} \quad (1.64)$$

The two above equations show that the function of the electric field is continuous, since on the surface of the cylinder $r = R$ substitution produces the same result.

On the surface of the sphere the numerical value of the electric field can readily be calculated:

$$E(R) = \frac{\rho}{2\varepsilon_0} R = \frac{10^{-6}}{2 \cdot 8,86 \cdot 10^{-12}} \cdot 0.1 = 5643 \frac{V}{m} \quad (1.65)$$

The potential function can be determined by integrating the electric field. First the internal region is integrated: The reference point of the potential will be the center.

$$U_{in}(r) = - \int_0^r E(r') dr' = - \int_0^r \frac{\rho}{2\varepsilon_0} r' dr' = - \frac{\rho}{2\varepsilon_0} \int_0^r r' dr' = - \frac{\rho}{2\varepsilon_0} \left[\frac{r'^2}{2} \right]_{r'=0}^{r'=r} = - \frac{\rho}{2\varepsilon_0} \left(\frac{r^2}{2} \right) = - \frac{\rho}{4\varepsilon_0} r^2 \quad (1.66)$$

$$U_{in}(r) = - \frac{\rho}{4\varepsilon_0} r^2 \quad (1.67)$$

The surface potential of the cylinder is the above function with $r = R$ substitution:

$$U_{in}(R) = - \frac{\rho R^2}{4\varepsilon_0} = - \frac{10^{-6} \cdot 10^{-2}}{4 \cdot 8,86 \cdot 10^{-12}} = -282V \quad (1.68)$$

Remember that this value should be added to the integral calculated next.

$$U_{out}(r) = U_{in}(R) + U_R^r = U_{in}(R) + \left(- \int_R^r E(r') dr' \right) \quad (1.69)$$

For simplicity reason only the integral in the parenthesis is transformed first:

$$U_R^r = - \int_R^r \frac{\rho}{2\varepsilon_0} \frac{R^2}{r'} dr' = - \frac{\rho R^2}{2\varepsilon_0} \int_R^r \frac{dr'}{r'} = - \frac{\rho R^2}{2\varepsilon_0} [\ln r']_{r'=R}^{r'=r} = - \frac{\rho R^2}{2\varepsilon_0} \ln \left(\frac{r}{R} \right) \quad (1.70)$$

Altogether:

$$U_{out}(r) = U_{in}(R) + U_R^r = - \frac{\rho R^2}{4\varepsilon_0} + \left(- \frac{\rho R^2}{2\varepsilon_0} \ln \left(\frac{r}{R} \right) \right) = - \frac{\rho R^2}{4\varepsilon_0} \left(1 + 2 \ln \left(\frac{r}{R} \right) \right) \quad (1.71)$$

The final result is:

$$U_{out}(r) = - \frac{\rho R^2}{4\varepsilon_0} \left(1 + 2 \ln \left(\frac{r}{R} \right) \right) \quad (1.72)$$

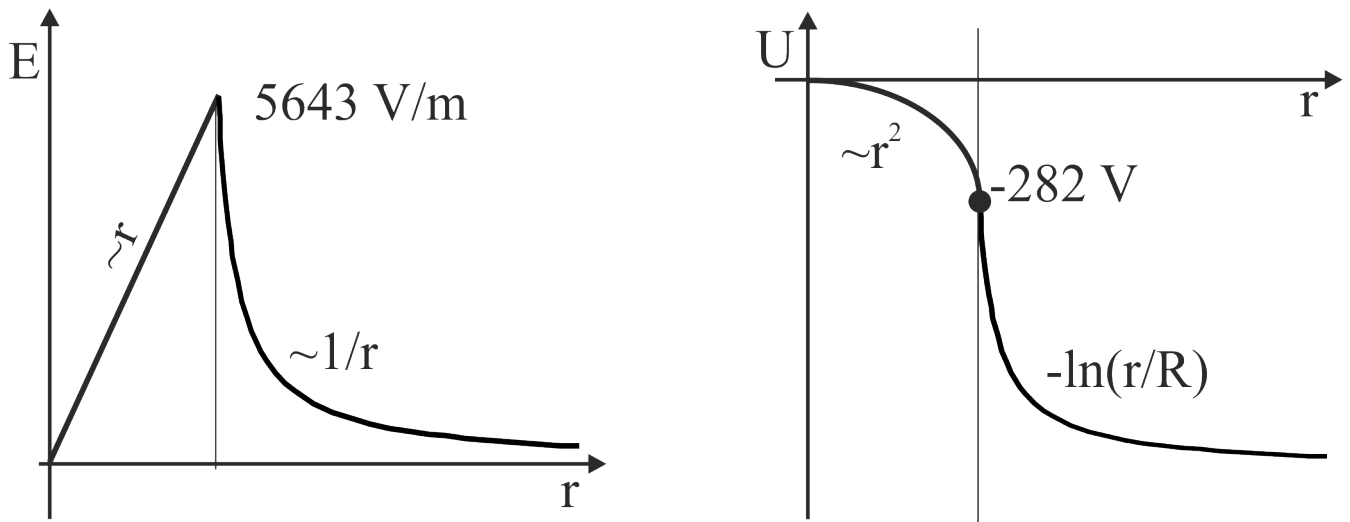


Figure 1.5: Cylinder with uniform charge density
 Electric field vs. radial position function Potential vs. radial position function

1.11 Infinite parallel plate with uniform surface charge density

Infinite metal plate contains $\omega = 10^{-8} \text{ As/m}^2$ surface charge density. Find the function of the electric field and the potential as the function of distance from the plate and sketch the result. The reference point of the potential should be the center. Calculate the value of the electric field on the surface of the metal plate.

$$\oint_S \mathbf{E}(\mathbf{r}) \cdot d\mathbf{A} = \frac{Q}{\varepsilon_0} \quad (1.73)$$

Gauss's law is used to solve the problem. Pick a virtual drum with base plate area A . Position the drum with rotational axis normal to the charged plate. The charged plate should cut the drum to two symmetrical parts.

$$E \cdot 2A = \frac{\omega A}{\varepsilon_0} \quad (1.74)$$

The absolute value electric field can be expressed:

$$E = \frac{\omega}{2\varepsilon_0} = \frac{10^{-8}}{2 \cdot 8.86 \cdot 10^{-12}} = 564 \frac{\text{V}}{\text{m}} \quad (1.75)$$

The result shows that the electric field is constant in the half space.

The direction of the electric field is opposite in the two half spaces. In contrast to the spherical and cylindrical structures where the radial distance is the position parameter, here a reference direction line will be used.

The potential function can be determined by integrating the electric field in the positive half space:

$$U(x) = - \int_0^x E(x') dx' = - \int_0^x \frac{\omega}{2\epsilon_0} dx' = - \frac{\omega}{2\epsilon_0} \int_0^x dx' = - \frac{\omega}{2\epsilon_0} [x']_{x'=0}^{x'=x} = - \frac{\omega}{2\epsilon_0} x \quad (1.76)$$

So briefly the potential function is as follows:

$$U(x) = - \frac{\omega}{2\epsilon_0} x \quad (1.77)$$

Obviously the potential function turns to its negative in the negative half space.

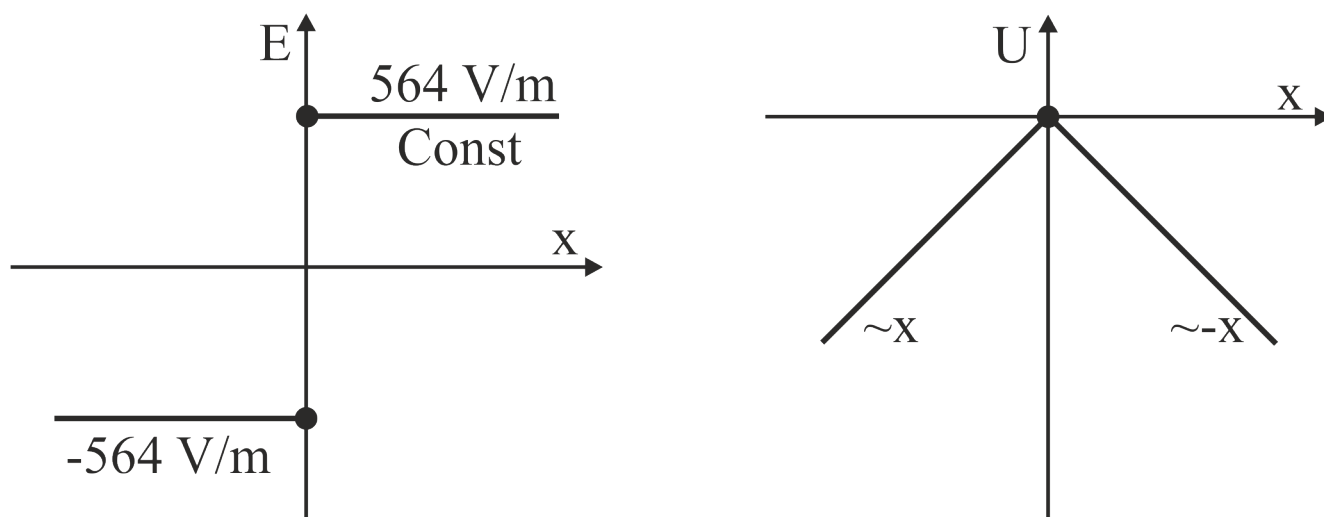


Figure 1.6: Infinite parallel plate with uniform surface charge density
 Electric field vs. radial position function Potential vs. radial position function

Note that the reference point of the potential could not be placed to the infinity because the infinite plate is not physically real object.

1.12 Capacitors

Capacitors consist of two plates to store charge. The overall contained charge is zero since the charges on the plates are opposite therefore the electric field is confined to the

inner volume of the capacitor. The capacitance is the ratio of the charge over the voltage generated between the plates. $C = \frac{Q}{U}$ Three different geometries will be treated below.

1.12.1/ Parallel plate capacitor

The parallel plate capacitor is made of two parallel metal plates facing each other with the active surface area A . The distance between the plates and the charge are denoted by d and Q , respectively.

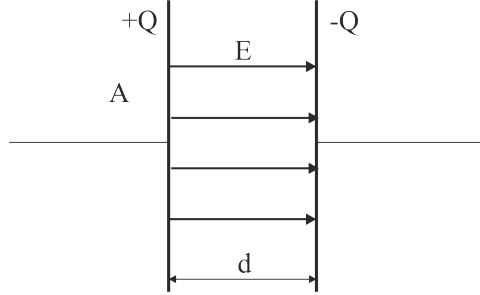


Figure 1.7: Parallel plate capacitor

There is homogeneous electric field between the plates, while out of the capacitor there is no electric field. Use the Gauss's law for a drum-like surface which surrounds one of the plates.

$$EA = \frac{Q}{\epsilon_0} \quad E = \frac{Q}{A\epsilon_0} \quad (1.78)$$

To find out the voltage between the plates does not need integration due to the homogeneous field. $U = d \cdot E$

$$U = \frac{d \cdot Q}{A\epsilon_0} \quad (1.79)$$

And capacitance can be expressed from here.

$$C = \epsilon_0 \frac{A}{d} \quad (1.80)$$

1.12.1 Cylindrical capacitor

The cylindrical capacitor is made of two coaxial metal cylinders. The inner and the outer radii as well as the length are denoted R_1, R_2 and l , respectively. The coaxial cable is the only practically used cylindrical capacitor.

Let us use the Gauss's law. A coaxial cylinder should be inflated from R_1 to R_2 .

$$\oint_S \mathbf{E}(\mathbf{r}) \cdot d\mathbf{A} = \frac{Q}{\epsilon_0} \quad (1.81)$$

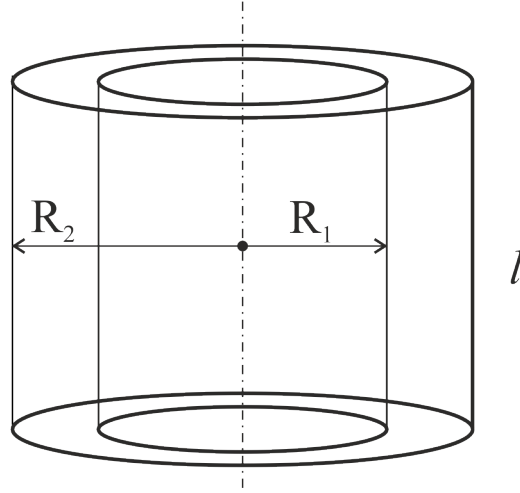


Figure 1.8: Cylindrical capacitor

$$E \cdot 2r\pi l = \frac{Q}{\varepsilon_0} \quad (1.82)$$

$$E = \frac{Q}{2\pi\varepsilon_0 l} \cdot \frac{1}{r} \quad (1.83)$$

To find out the voltage the following integral should be evaluated:

$$U = - \int_{R_2}^{R_1} \frac{Q}{2\pi\varepsilon_0 l} \cdot \frac{1}{r} dr = \frac{Q}{2\pi\varepsilon_0 l} \int_{R_1}^{R_2} \frac{1}{r} dr = \frac{Q}{2\pi\varepsilon_0 l} \ln\left(\frac{R_2}{R_1}\right) \quad (1.84)$$

The capacitance can be expressed from here.

$$C = \frac{Q}{U} = \frac{2\pi\varepsilon_0 l}{\ln(R_2/R_1)} \quad (1.85)$$

The above formula shows the obvious fact that the capacitance is proportional to the length of the structure. Because of this the capacitance of one meter coaxial cable is used mostly. This is denoted with c and measured in F/m units. Most coaxial cables represent some $10 \text{ pF}/m$ value.

$$c = \frac{Q}{U} = \frac{2\pi\varepsilon_0}{\ln(R_2/R_1)} \quad (1.86)$$

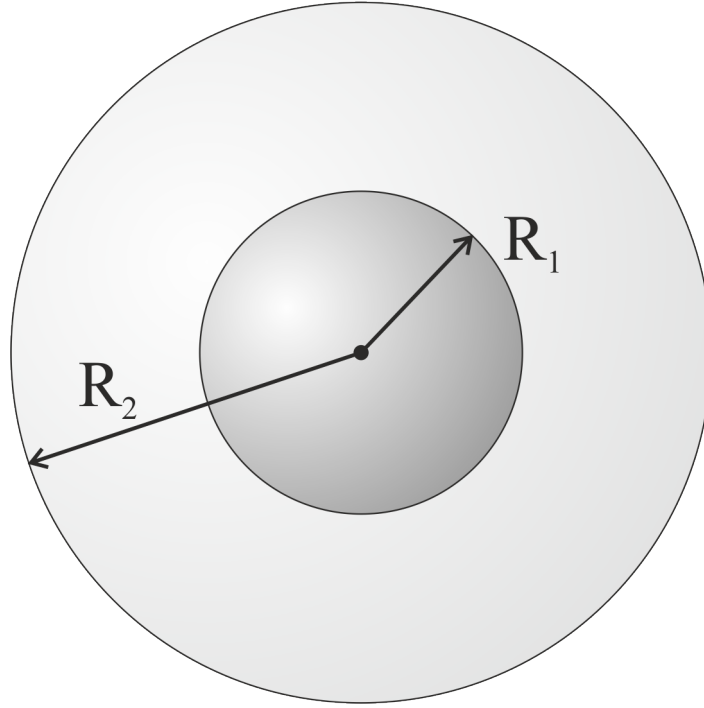


Figure 1.9: Spherical capacitor

1.12.2 Spherical capacitor

The spherical capacitor is made of two concentric metal spheres. The inner and the outer radii as well as the charge are denoted R_1 , R_2 and Q , respectively.

Let us use the Gauss's law. A concentric sphere should be inflated from R_1 to R_2 .

$$\oint_S \mathbf{E}(\mathbf{r}) \cdot d\mathbf{A} = \frac{Q}{\varepsilon_0} \quad (1.87)$$

$$E \cdot 4r^2\pi = \frac{Q}{\varepsilon_0} \quad (1.88)$$

$$E = \frac{Q}{4\pi\varepsilon_0} \cdot \frac{1}{r^2} \quad (1.89)$$

To find out the voltage the following integral should be evaluated:

$$U = - \int_{R_2}^{R_1} \frac{Q}{4\pi\varepsilon_0} \cdot \frac{1}{r^2} dr = \frac{Q}{4\pi\varepsilon_0} \int_{R_2}^{R_1} \left(-\frac{1}{r^2}\right) dr = \frac{Q}{4\pi\varepsilon_0} \left[\frac{1}{r} \right]_{r=R_2}^{r=R_1} = \frac{Q}{4\pi\varepsilon_0} \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (1.90)$$

The capacitance can be expressed from here.

$$C = \frac{Q}{U} = \frac{4\pi\epsilon_0}{\frac{1}{R_1} - \frac{1}{R_2}} \quad (1.91)$$

1.13 Principle of superposition

The Gauss's law can only be used effectively in the three symmetry classes mentioned earlier. If the charge arrangement does not belong to any of those classes the principle of superposition is the only choice. This time the charge arrangement is virtually broken to little pieces and the electric fields of these little pieces are superimposed like point charges.

- Find the electric field of a finite long charged filament in the equatorial plane as the function of distance from the filament. The linear charge density is denoted σ .

Since the filament is not infinite long Gauss's law can not be used effectively. The charged filament is divided to little infinitesimal pieces and the electric fields of such pieces are added together. Due to symmetry reasons only the normal components of the electric field are integrated since the parallel components cancel out by pairs. The mathematical deduction of the final formula follows below without close commenting to the transformations. For the definition of the notations refer the figure below:

The infinitesimal contribution of the electric field is calculated as a point charge.

$$dE^* = \frac{dQ}{4\pi\epsilon_0} \cdot \frac{1}{r^2} \quad dQ = \frac{rd\phi}{\cos\varphi} \sigma \quad r = \frac{R}{\cos\varphi} \quad (1.92)$$

The infinitesimal charge is contained by the infinitesimal angle.

$$dE^* = \frac{rd\phi}{\cos\varphi} \sigma \frac{1}{4\pi\epsilon_0} \cdot \frac{1}{r^2} = \frac{\sigma}{4\pi\epsilon_0} \frac{1}{r \cos\varphi} d\phi = \frac{\sigma}{4\pi\epsilon_0} \frac{1}{R} d\phi \quad (1.93)$$

The electric field of the point charge is projected to the perpendicular direction. The parallel direction components cancel out by symmetric pairs.

$$dE = dE^* \cos\phi \quad (1.94)$$

$$dE = \frac{\sigma}{4\pi\epsilon_0} \frac{1}{R} \cos\phi \cdot d\phi \quad (1.95)$$

The integration is carried out in α half visual angle.

$$E = \int_{-\alpha}^{\alpha} dE = \int_{-\alpha}^{\alpha} \frac{\sigma}{4\pi\epsilon_0} \frac{1}{R} \cos\phi \cdot d\phi = \frac{\sigma}{4\pi\epsilon_0} \frac{1}{R} \int_{-\alpha}^{\alpha} \cos\phi \cdot d\phi = \frac{\sigma}{4\pi\epsilon_0} \frac{1}{R} [\sin\phi]_{\phi=-\alpha}^{\phi=\alpha} = \frac{\sigma}{2R\pi\epsilon_0} \sin\alpha \quad (1.96)$$

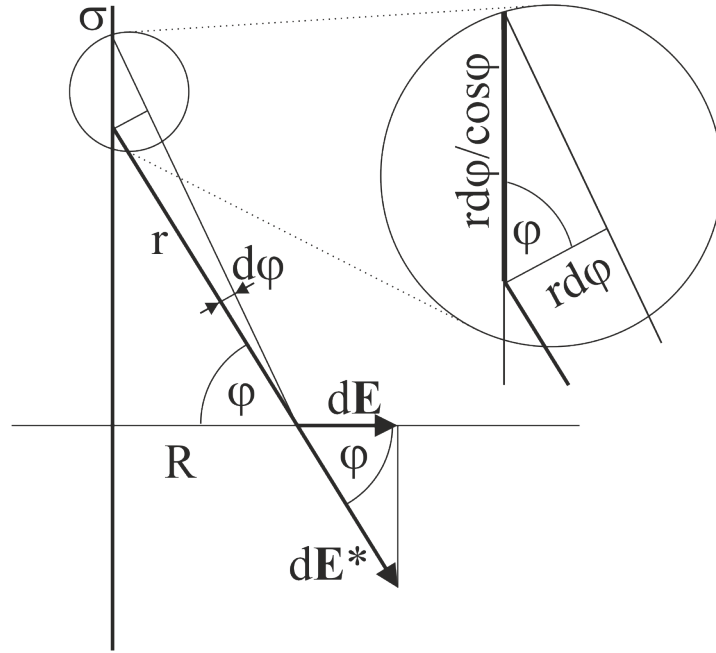


Figure 1.10: Principle of superposition

The result of the superposition is the formula below which could not have been attained with Gauss's law.

$$E = \frac{\sigma}{2R\pi\epsilon_0} \sin \alpha \quad (1.97)$$

If α approaches ninety degrees the filaments tend to the infinity when using Gauss's law is an option.

$$E_\infty = \lim_{\alpha \rightarrow \frac{\pi}{2}} E = \frac{\sigma}{2R\pi\epsilon_0} \quad (1.98)$$

Using Gauss's law the above result can be reached far easier for the infinite long filament.

$$E_\infty \cdot 2R\pi \cdot l = \frac{\sigma l}{\epsilon_0} \quad (1.99)$$

$$E_\infty = \frac{\sigma}{2R\pi\epsilon_0} \quad (1.100)$$

The results are in perfect match. However the point is that superposition principle can be used in full generality, but it is far more meticulous and tedious than using Gauss's law if that is possible.

Chapter 2

Dielectric materials - György Hárs

Insulators or in other words dielectric materials will be discussed in this chapter. In chapter 1 only metal electrodes and immobile charges are the sources of the electric field. In contrast to metals, insulating materials are lacking of mobile electron plasma therefore the electric field can penetrate the insulators. In terms of phenomenology the insulating material is polarized which means that the material as a whole will become an electric dipole. In terms of microphysical explanation, the overall effects of huge number of elementary dipoles will create the external dipole effect. The elementary dipoles are either generated or oriented by the external electric field.

2.1 The electric dipole

Consider a pair of opposite point charges $(+q, -q)$. Initiate the vector of separation (\mathbf{s}) from the negative to the positive point charge. The following product defines the electric dipole moment:

$$\mathbf{p} = q\mathbf{s} \quad [Asm] \quad (2.1)$$

In order to generate a point-like dipole the definition is completed with a limit transition. Accordingly the absolute value of the displacement vector shrinks to zero while the charge tends to the infinity such a way that the product is a constant vector. The point-like dipole is a useful model when the distance of the charges is far smaller than the corresponding geometry, for example if a dipole molecule is located in the proximity of centimeter size electrodes.

Force couple is exerted to the dipole by homogeneous electric field. The torque (\mathbf{M}) generated turns the dipole parallel to the electric field.

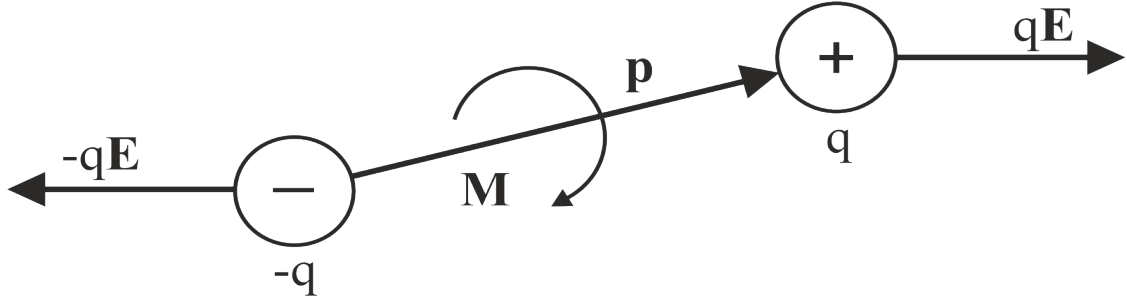


Figure 2.1: Dipole turns parallel to the \mathbf{E} field

$$\mathbf{M} = \mathbf{p} \times \mathbf{E} \quad \left[A \text{sm} \cdot \frac{V}{m} = Nm \right] \quad (2.2)$$

The dipole moment turns into the direction of the electric field spontaneously and stays there. Having reached this position, the least amount of potential energy is stored by the dipole. Obviously the most amount of potential energy stored is just in the opposite position. Let us find out the work needed to turn the dipole from the deepest position to the highest energy.

$$W = \int_0^\pi M d\phi = \int_0^\pi pE \sin \phi \cdot d\phi = pE \int_0^\pi \sin \phi \cdot d\phi = -pE [\cos \phi]_0^\pi = 2pE \quad (2.3)$$

According to this result the potential energy of the dipole is as follows:

$$E_{pot} = -\mathbf{p} \cdot \mathbf{E} \quad (2.4)$$

This formula provides the deepest energy at parallel spontaneous position and the highest at anti-parallel position. The zero potential energy is at ninety degrees. The difference between the highest and lowest is just the work needed to turn it around.

2.2 Polarization

Take a plate capacitor and fill its volume with a dielectric material. The experiment shows that the capacitance increased relative to the empty case. The explanation behind is the polarization of the dielectric material.

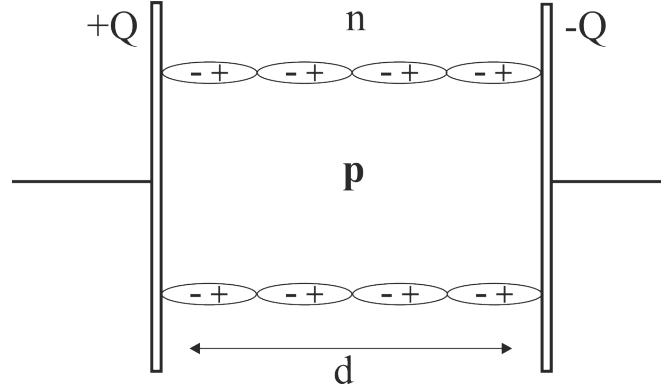


Figure 2.2: Plate capacitor with dipole chain

Due to the electric field of the metal plates the atomic dipoles have been arranged like the chains as the figure shows above. Inside the dielectric material the electric effect of dipoles cancel out since in any macroscopic volume equal number of positive and negative charges is present. The exceptions are the two sides of the dielectric material where the uncompensated polarization surface charge densities reside. These uncompensated polarization surface charges are opposite in polarity relative to the adjacent metal plates. This way the effective total charge is reduced, thus voltage of the capacitor diminished and ultimately the capacitance is increased.

Assume that the insulating material contains n pieces of dipoles per unit volume ($1/m^3$). The surface area and the separation of the plate capacitor is denoted A and d respectively. The total number of dipoles (N) is as follows:

$$N = Adn \quad (2.5)$$

The dipoles are located in chains between the metal plates. The number of dipoles in such a chain is the ratio of the distance between the plates (d) and the separation of the dipoles (s). The total number of dipoles (N) can be expressed if the length of the chain is multiplied with the number of chains (c) present in the material:

$$N = \frac{d}{s}c \quad (2.6)$$

Let us combine these latter two equations:

$$Adn = \frac{d}{s}c \quad Ans = c \quad (2.7)$$

The separation (s) can be expressed as ratio of the dipole moment (p) and the charge of the dipole (q). (In present discussion the absolute values of the quantities are denoted without vector notation.)

$$s = \frac{p}{q} \quad (2.8)$$

So the number of chains can be expressed:

$$c = An \frac{p}{q} \quad (2.9)$$

The total polarization surface charge is the product of the number of chains and the uncompensated opposite charge at the end of each dipole chain.

$$Q_p = c(-q) = -Anp \quad (2.10)$$

Finally the polarization surface charge density (ω_p) needs to be expressed:

$$\omega_p = \frac{Q_p}{A} = -np = -P \quad \left[\frac{As}{m^2} \right] \quad (2.11)$$

Here we introduced the vector of the “polarization (\mathbf{P})”. The sources of this vector are the opposite of the polarization charges. The negative polarity comes from the definition of the electric dipole which points from to minus to the plus in contrast to the direction of the electric field. (Without vector notation the absolute value is meant).

An additional result can also be concluded. The density of dipoles gives rise to the polarization. The formula is valid in three dimensions too.

$$P = np \quad \mathbf{P} = np \quad (2.12)$$

2.3 Dielectric displacement

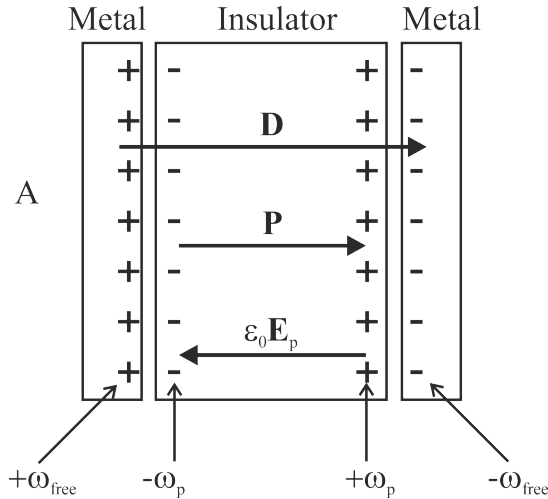


Figure 2.3: Plate capacitor with polarized dielectric material

The metal plates of the capacitor contain what are called the “free charges”. The free charges are mobile and can be conducted away by means of a wire. In contrast to this the polarization charges are immobile.

In chapter 1 the homogeneous electric field in plate capacitor has been expressed, provided the free surface charge densities $(+\omega_{\text{free}}, -\omega_{\text{free}})$ are located on the metal plates.

$$E_{\text{free}} = \frac{\omega_{\text{free}}}{\varepsilon_0} \quad \varepsilon_0 E_{\text{free}} = \omega_{\text{free}} \quad (2.13)$$

Here we introduce the vector of the “dielectric displacement (**D**)”. The sources of this vector are the free mobile charges. (Without vector notation the absolute value is meant).

$$\varepsilon_0 E_{\text{free}} = D = \omega_{\text{free}} \left[\frac{As}{m^2} \right] \quad (2.14)$$

The polarization surface charges $(+\omega_p, -\omega_p)$ perform the same way but in opposite direction:

$$E_p = \frac{\omega_p}{\varepsilon_0} \quad \varepsilon_0 E_p = \omega_p \quad (2.15)$$

$$\varepsilon_0 E_p = -P = \omega_p \left[\frac{As}{m^2} \right] \quad (2.16)$$

The total surface charge density is the sum of the free and the polarization charges:

$$\omega_{\text{tot}} = \omega_{\text{free}} + \omega_p \quad (2.17)$$

The total charge density is the source of the resulting electric field in the capacitor:

$$\varepsilon_0 E = D - P \quad (2.18)$$

$$D = \varepsilon_0 E + P \quad (2.19)$$

The last formula has been deduced for one dimensional case. The parameters show up here as they were real numbers. However the result is true in full generality in three dimensions with vectors as well.

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} \quad (2.20)$$

Note the important fact that the electric field and so the intensity of forces are always reduced in presence of dielectric material.

2.4 Electric permittivity (dielectric constant)

Experiments show that the polarization of some isotropic material is the monotonous function of the external electric field. At external fields of higher intensity the insulating material is gradually saturated. At relatively low levels the function can be considered linear. Our next discussion is confined to the linear range. This case the proportionality is holding between the polarization and the external electric field. In order to make an equation out of the proportionality a coefficient (χ) is introduced.

$$\mathbf{P} = \chi \varepsilon_0 \mathbf{E} \quad (2.21)$$

The coefficient is the permittivity of vacuum (ε_0) and the electric susceptibility (χ).

Substitute this equation to the former expression of \mathbf{D} vector:

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \chi \varepsilon_0 \mathbf{E} = \varepsilon_0(1 + \chi) \mathbf{E} = \varepsilon_0 \varepsilon_r \mathbf{E} \quad (2.22)$$

Here the relative permittivity (ε_r) has been introduced:

$$1 + \chi = \varepsilon_r \quad (2.23)$$

Typical values of the relative permittivity are up to five or so. Very high numbers are technically impossible.

Finally the result to be remembered is as follows:

$$\mathbf{D} = \varepsilon_0 \varepsilon_r \mathbf{E} \quad \left[\frac{As}{m^2} \right] \quad (2.24)$$

2.5 Gauss's law and the dielectric material

In this section the vector calculus will be used at somewhat higher level.

The divergence operation (*div*) generates a scalar field which represents the sources of some vector field.

$$\mathbf{V}(\mathbf{r}) = V_x(x, y, z)\mathbf{i} + V_y(x, y, z)\mathbf{j} + V_z(x, y, z)\mathbf{k} \quad (2.25)$$

$$div \mathbf{V}(\mathbf{r}) = \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} + \frac{\partial V_z}{\partial z} \quad (2.26)$$

The Gauss Ostrogradsky theorem integrates the divergence to a volume as follows:

$$\oint_S \mathbf{V}(\mathbf{r}) d\mathbf{A} = \oint_V (div \mathbf{V}) dV \quad (2.27)$$

Let us generate the divergence of the equation discussed earlier in this chapter:

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} \quad (2.28)$$

$$\text{div} \mathbf{D} = \text{div}(\varepsilon_0 \mathbf{E}) + \text{div} \mathbf{P} \quad (2.29)$$

The sources or in other words the divergences are the corresponding volume charge densities (ρ_{free} , ρ_{tot} , ρ_p) in general. Earlier in this chapter the surface charge densities have been discussed in details for the case of the plate capacitor. Accordingly the following relations are plausible:

$$\text{div} \mathbf{D} = \rho_{\text{free}} \quad \text{div}(\varepsilon_0 \mathbf{E}) = \rho_{\text{tot}} \quad \text{div} \mathbf{P} = -\rho_p \quad (2.30)$$

Gauss Ostrogradsky theorem generates integral form from the relations above:

$$\oint_S \mathbf{D} d\mathbf{A} = Q_{\text{free}} \quad \oint_S (\varepsilon_0 \mathbf{E}) d\mathbf{A} = Q_{\text{tot}} \quad \oint_S \mathbf{P} d\mathbf{A} = -Q_p \quad (2.31)$$

The left hand side integral is the well-known form of Gauss's law with \mathbf{D} vector. This expresses that the flux of \mathbf{D} vector on a closed surface (S) equals the amount of the contained free charges. The integral in the middle expresses that the flux of $\varepsilon_0 \mathbf{E}$ vector equals the total amount of any contained charges. Finally the right hand side states that the flux of the polarization \mathbf{P} vector equals the opposite of the polarization charges contained by the S surface.

2.6 Inhomogeneous dielectric materials

Consider two different dielectric materials with plane surface. The plane surfaces are connected thus creating an interface between the insulators. This structure is subjected to the experimentation.

First the \mathbf{D} field is studied.

The interface is contained by a symmetrical disc-like drum with the base area A . The upper and lower surface vectors are \mathbf{A}_1 and \mathbf{A}_2 respectively.

$$\mathbf{A}_1 = -\mathbf{A}_2 \quad |\mathbf{A}_1| = |\mathbf{A}_2| = A \quad (2.32)$$

The volume does not contain free charges therefore the flux of the \mathbf{D} vector is zero.

$$\oint_S \mathbf{D} d\mathbf{A} = \mathbf{D}_1 \mathbf{A}_1 + \mathbf{D}_2 \mathbf{A}_2 = 0 \quad (2.33)$$

$$\mathbf{D}_1 \mathbf{A}_2 = \mathbf{D}_2 \mathbf{A}_2 \quad (2.34)$$

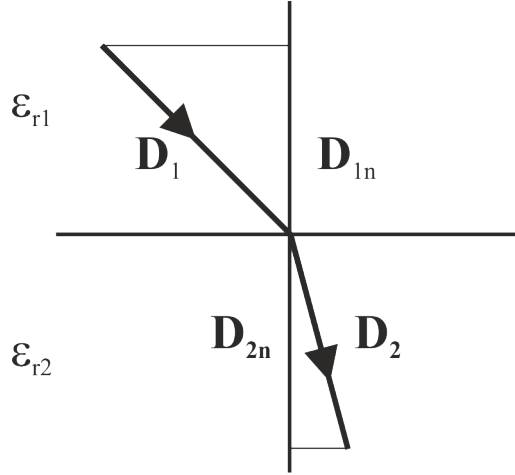


Figure 2.4: \mathbf{D} field at the interface of different dielectric materials

The operation of dot product contains the projection of the \mathbf{D} vectors to the direction of \mathbf{A}_2 vector which is the normal direction to the surface. The subscript n means the absolute value of the normal direction component.

$$D_{1n}A_2 = D_{2n}A_2 \quad (2.35)$$

Once we are among real numbers the surface area cancels out readily.

$$D_{1n} = D_{2n} \quad (2.36)$$

According to this result the normal component of \mathbf{D} vector is continuous on the interface of dielectric materials.

Secondly the electric field \mathbf{E} is the subject of analysis.

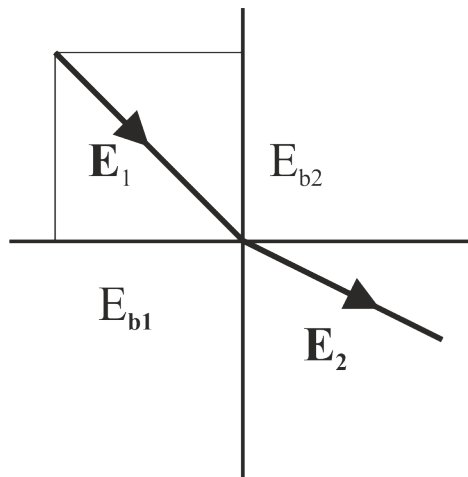


Figure 2.5: \mathbf{E} field at the interface of different dielectric materials

The interface is surrounded by a very narrow rectangle-like loop with sections parallel and normal to the surface. The parallel sections of the loop are \mathbf{s} and $-\mathbf{s}$ vectors. The normal direction sections are ignored due to the infinitesimal size. The closed loop integral of the \mathbf{E} in static electric field is zero.

$$\oint_g \mathbf{E} d\mathbf{r} = \mathbf{sE}_1 + (-\mathbf{s})\mathbf{E}_2 = 0 \quad (2.37)$$

$$\mathbf{sE}_1 = \mathbf{sE}_2 \quad (2.38)$$

The operation of dot product contains the projection of the \mathbf{E} vectors to the direction of \mathbf{s} vector which is the tangential direction to the surface. The subscript t means the absolute value of the tangential direction component.

$$sE_{1t} = sE_{2t} \quad (2.39)$$

Once we are among real numbers the length of the tangential section cancels out readily.

$$E_{1t} = E_{2t} \quad (2.40)$$

According to this result the tangential component of the \mathbf{E} vector is continuous on the interface of dielectric materials.

2.7 Demonstration examples

2.7.1

A metal sphere with radius ($R_1 = 10\text{cm}$) contains free charges ($Q_{free} = 10^{-8}\text{As}$). The metal sphere is surrounded by an insulating layer ($\epsilon_r = 3$) up to the radius ($R_2 = 15\text{cm}$). Find and sketch the radial dependence of D, E and P vectors. Determine the numerical peak values in the break points and find the amount of the polarization charge.

The first parameter to deal with is the D vector because the normal component is continuous on the interface of dielectric materials. Let us use the Gauss's law.

$$\oint_S \mathbf{D} d\mathbf{A} = Q_{free} \quad (2.41)$$

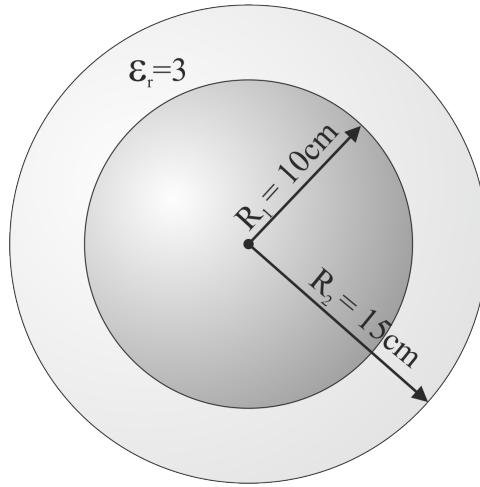


Figure 2.6: Metal sphere surrounded by insulating layer

Inside the metal sphere all the parameters are zero only out of the metal sphere is of interest.

$$4r^2\pi \cdot D = Q_{free} \quad (2.42)$$

$$D = \frac{Q_{free}}{4\pi} \cdot \frac{1}{r^2} \quad (2.43)$$

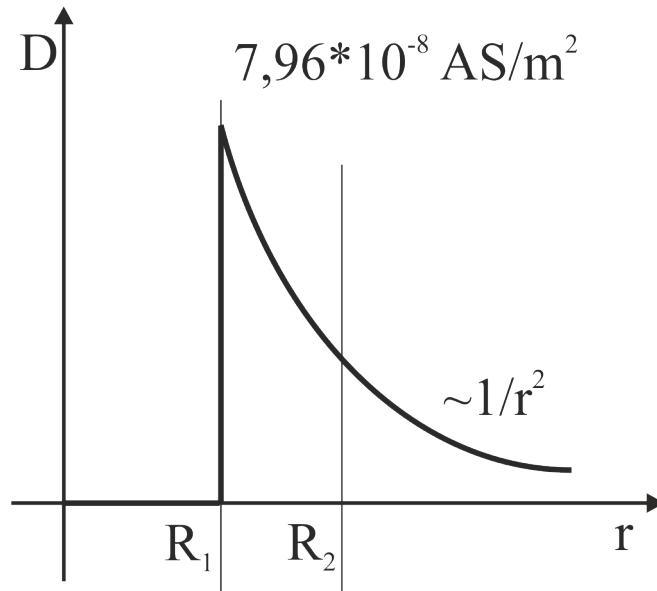


Figure 2.7: The absolute value of D vs. radial position function

The peak value at the brake point results once $r = R_1$ is substituted.

$$D(R_1) = \frac{Q_{free}}{4\pi} \cdot \frac{1}{R_1^2} = \frac{10^{-8}}{4\pi} 100 = 7.96 \cdot 10^{-8} \frac{As}{m^2} \quad (2.44)$$

The $\varepsilon_0 E$ field is identical with the D function out of the insulator. In the insulator however the $\varepsilon_0 E$ function is reduced to one third, according to $\varepsilon_r = 3$ value.

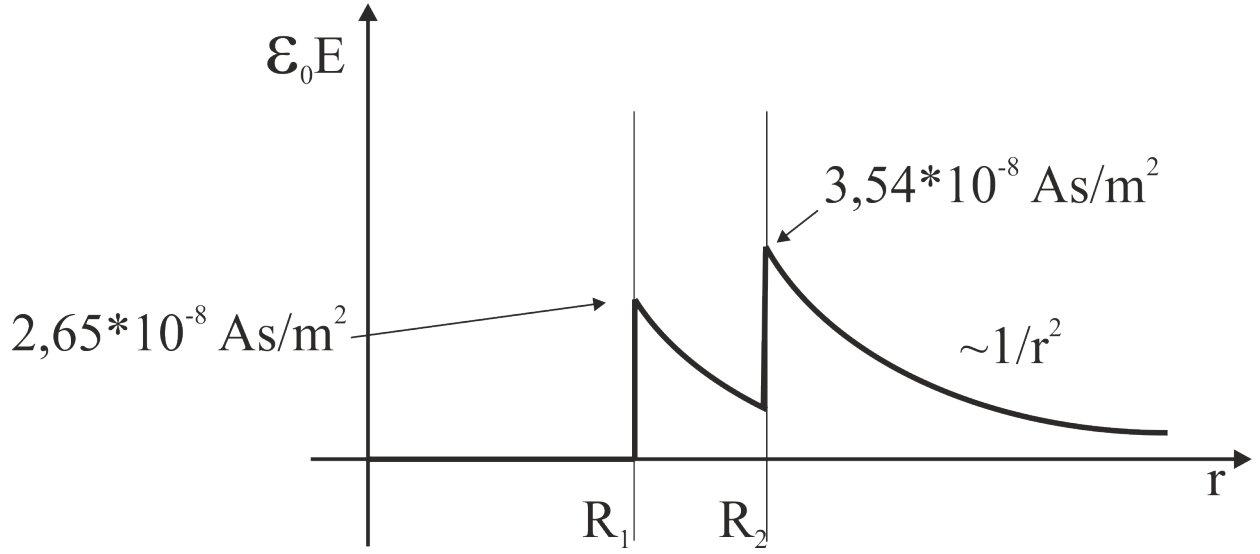


Figure 2.8: The absolute value of $\varepsilon_0 E$ vs. radial position function

The $\varepsilon_0 E$ functions in the insulator and out of the structure are as follows:

$$\varepsilon_0 E_{in} = \frac{Q_{free}}{4\pi\varepsilon_r} \cdot \frac{1}{r^2} \quad \varepsilon_0 E_{out} = \frac{Q_{free}}{4\pi} \cdot \frac{1}{r^2} \quad (2.45)$$

The brake point peak values of $\varepsilon_0 E$ function are as follows:

$$\varepsilon_0 E_{in}(R_1) = \frac{Q_{free}}{4\pi\varepsilon_r} \cdot \frac{1}{R_1^2} = \frac{10^{-8}}{4\pi \cdot 3} \cdot \frac{1}{0.1^2} = 2.65 \cdot 10^{-8} \frac{As}{m^2} \quad (2.46)$$

$$\varepsilon_0 E_{out}(R_2) = \frac{Q_{free}}{4\pi} \cdot \frac{1}{R_2^2} = \frac{10^{-8}}{4\pi} \cdot \frac{1}{0.15^2} = 3.54 \cdot 10^{-8} \frac{As}{m^2} \quad (2.47)$$

The corresponding electric fields are:

$$E_{in}(R_1) = \frac{2.65 \cdot 10^{-8}}{8.86 \cdot 10^{-12}} = 3000 \frac{V}{m} \quad E_{out}(R_2) = \frac{3.54 \cdot 10^{-8}}{8.86 \cdot 10^{-12}} = 4000 \frac{V}{m} \quad (2.48)$$

The radial function of the P vector is zero except for the insulating material. In the insulating material this is as follows:

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} \quad (2.49)$$

$$P = D - \varepsilon_0 E_{in} \quad (2.50)$$

$$P = \frac{Q_{free}}{4\pi} \cdot \frac{1}{r^2} - \frac{Q_{free}}{4\pi\varepsilon_r} \cdot \frac{1}{r^2} = \left(1 - \frac{1}{\varepsilon_r}\right) \frac{Q_{free}}{4\pi} \cdot \frac{1}{r^2} \quad (2.51)$$

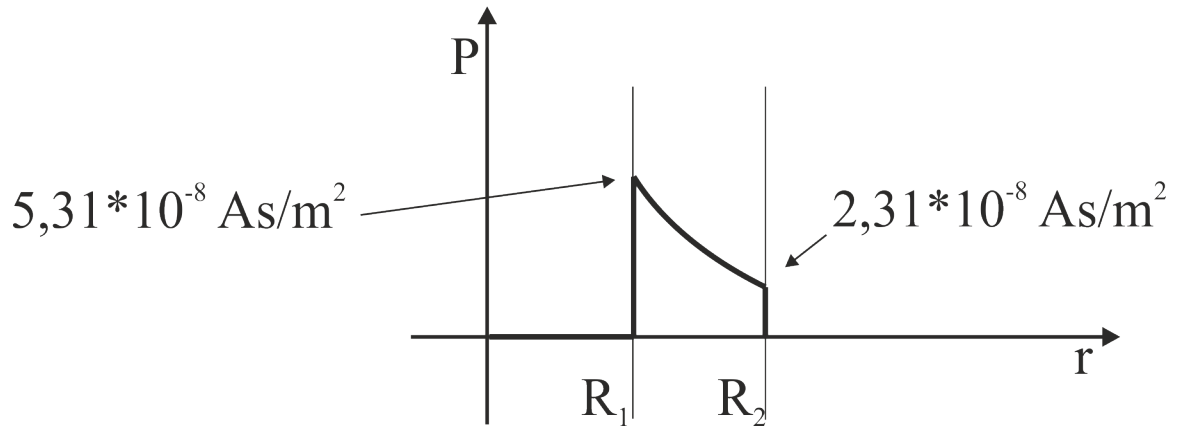


Figure 2.9: The absolute value of P vs. radial position function

Let us determine the peak values in the break points:

$$P(R_1) = \left(1 - \frac{1}{\varepsilon_r}\right) \frac{Q_{free}}{4\pi} \cdot \frac{1}{R_1^2} = \left(1 - \frac{1}{3}\right) \frac{10^{-8}}{4\pi} \frac{1}{0.1^2} = 5.31 \cdot 10^{-8} \frac{As}{m^2} \quad (2.52)$$

$$P(R_2) = \left(1 - \frac{1}{\varepsilon_r}\right) \frac{Q_{free}}{4\pi} \cdot \frac{1}{R_2^2} = \left(1 - \frac{1}{3}\right) \frac{10^{-8}}{4\pi} \frac{1}{0.15^2} = 2.36 \cdot 10^{-8} \frac{As}{m^2} \quad (2.53)$$

The amount of the polarization charge can be calculated:

$$\oint_S \mathbf{P} d\mathbf{A} = -Q_p \quad (2.54)$$

$$4r^2\pi \cdot \left(1 - \frac{1}{\varepsilon_r}\right) \frac{Q_{free}}{4\pi} \cdot \frac{1}{r^2} = -Q_p \quad (2.55)$$

$$Q_p = -\left(1 - \frac{1}{\varepsilon_r}\right) Q_{free} = -\frac{2}{3} 10^{-8} = -6.66 \cdot 10^{-9} As \quad (2.56)$$

2.7.2

Study the results of the previous demonstration example in that hypothetic case (Case 1) if the relative permittivity tends to the infinity. Compare the results with the case (Case 2) when the dielectric material would be replaced with metal.

It is interesting to observe that the E field is identical in both cases. The sources of the electric field are the total charges so the free and the polarization charges both count. The D field is different since in Case 1 the function did not change but in Case 2 it vanished between the radii. This happened because the sources of the D field are solely the free charges, so in Case 1 it did not change while in Case 2 it did change due to the free charges generated by metal. The P and the D field compensate each other so E field has been reduced to zero between the radii in Case1. In Case 2 P vector is obviously zero in absence of dielectric material.

2.8 Energy relations

Any electrostatic charge arrangement represents potential energy. This energy equals the amount of work needed to create the arrangement.

2.8.1 Energy stored in the capacitor

Consider a capacitor without charges initially. Carry an infinitesimal amount of dQ charge from one plate to the other. Therefore voltage (dQ/C) will appear between the plates. The next packet of dQ charge needs to be carried against the electric field generated by the previous packets. This way the voltage on the capacitor and the infinitesimal amounts of works will increase linearly. The triangle under the graph can represent the work done.

The total amount of work can be calculated by integration of those infinitesimal contributions.

$$E_{pot} = W = \int_0^Q U(Q') dQ' = \int_0^Q \frac{Q'}{C} dQ' = \frac{1}{C} \int_0^Q Q' dQ' = \frac{1}{C} \left[\frac{Q'^2}{2} \right]_{Q'=0}^{Q'=Q} = \frac{1}{2} \frac{Q^2}{C} \quad (2.57)$$

Relative permittivity tends to the infinity.
Case 1.

$$D = \frac{Q_{free}}{4\pi} \cdot \frac{1}{r^2} \quad R_1 \leq r$$

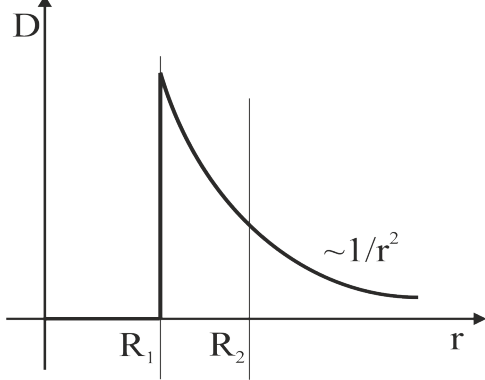


Fig. 2.10

The D field vs. radial position function

Insulating layer is replaced with metal. Case 2.

$$D = \frac{Q_{free}}{4\pi} \cdot \frac{1}{r^2} \quad R_2 \leq r$$

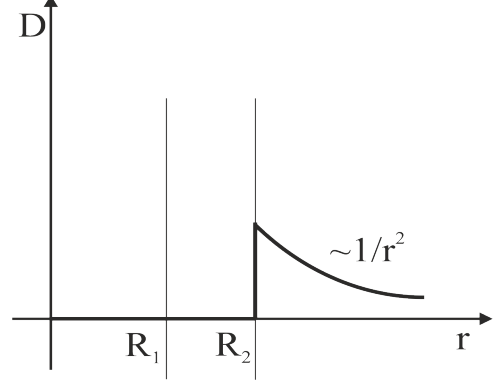


Fig. 2.11

The D field vs. radial position function

$$\varepsilon_0 E_{out} = \frac{Q_{free}}{4\pi} \cdot \frac{1}{r^2} \quad R_2 \leq r$$

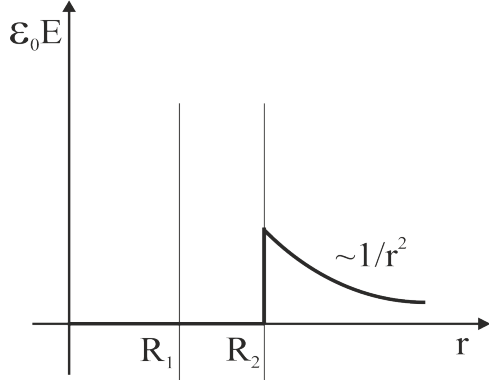


Fig. 2.12

The $\varepsilon_0 E$ field vs. radial position function

$$\varepsilon_0 E_{out} = \frac{Q_{free}}{4\pi} \cdot \frac{1}{r^2} \quad R_2 \leq r$$

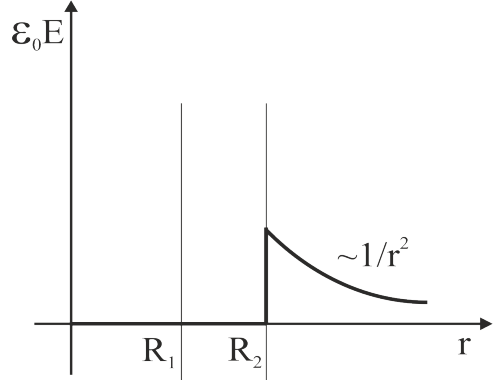


Fig. 2.13

The $\varepsilon_0 E$ field vs. radial position function

$$P = \frac{Q_{free}}{4\pi} \cdot \frac{1}{r^2} \quad R_1 \leq r \leq R_2$$

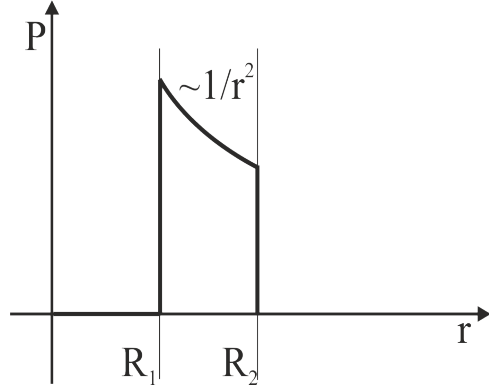


Fig. 2.14

The P field vs. radial position function

$$P = 0$$

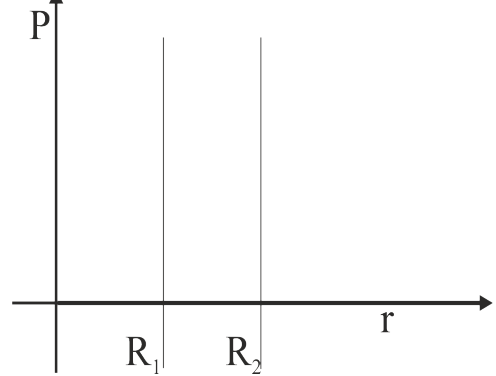


Fig. 2.15

The P field vs. radial position function

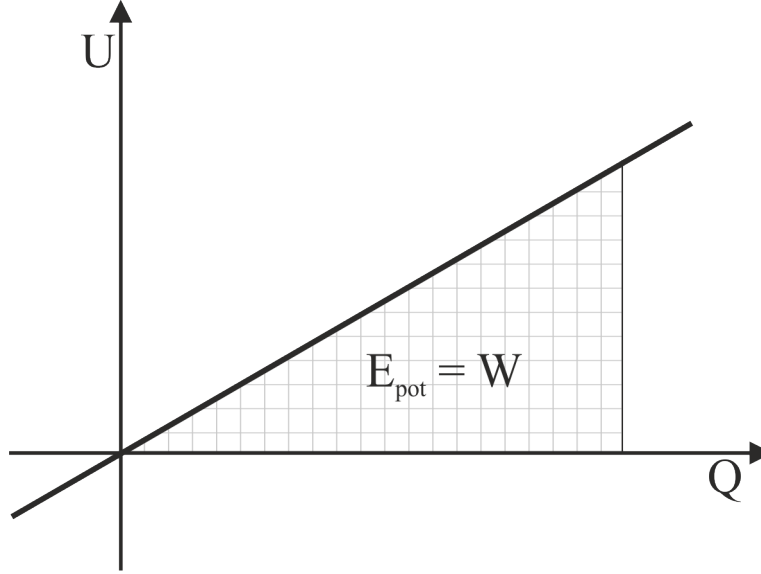


Figure 2.16: Voltage vs. charge function

The fundamental formula can be combined into the result. $Q = CU$

$$E_{pot} = \frac{1}{2} \frac{Q^2}{C} = \frac{1}{2} \frac{(CU)^2}{C} = \frac{1}{2} CU^2 \quad (2.58)$$

The practical cases use the last formula since the voltage is the known parameter mostly.

2.8.2 Electrostatic energy density

A plate capacitor is studied. The following pieces of information are at disposal:

$$U = Ed \quad C = \varepsilon_0 \varepsilon_r \frac{A}{d} \quad E_{pot} = \frac{1}{2} CU^2 \quad (2.59)$$

The notations are as defined earlier. Let us substitute to the final formula:

$$E_{pot} = \frac{1}{2} CU^2 = \frac{1}{2} \varepsilon_0 \varepsilon_r \frac{A}{d} (Ed)^2 = \frac{1}{2} \varepsilon_0 \varepsilon_r E^2 (Ad) \quad (2.60)$$

In the last formula the volume of the plate capacitor emerges. The energy density (e_{pot}) can be calculated as follows:

$$e_{pot} = \frac{E_{pot}}{Ad} = \frac{1}{2} \varepsilon_0 \varepsilon_r E^2 = \frac{1}{2} E(\varepsilon_0 \varepsilon_r E) = \frac{1}{2} ED \quad \left[\frac{J}{m^3} \right] \quad (2.61)$$

This result is also true in full generality in isotropic insulators. This time the dot product of the vectors is used.

$$e_{pot} = \frac{1}{2} \mathbf{E} \mathbf{D} \quad \left[\frac{J}{m^3} \right] \quad (2.62)$$

2.8.2 Principle of the virtual work

Electrostatic forces can be determined with the principle of the virtual work, provided the potential energy of a charge arrangement can be expressed as the function of some kind of position coordinate. This time the derivative of the potential energy results the intensity of force. The tedious integration of Coulomb's law can be replaced with the calculation of the potential energy, which is far easier task in most cases.

Demonstration example

Find the pressure exerted to the dielectric material between the two plates of a charged and disconnected plate capacitor.

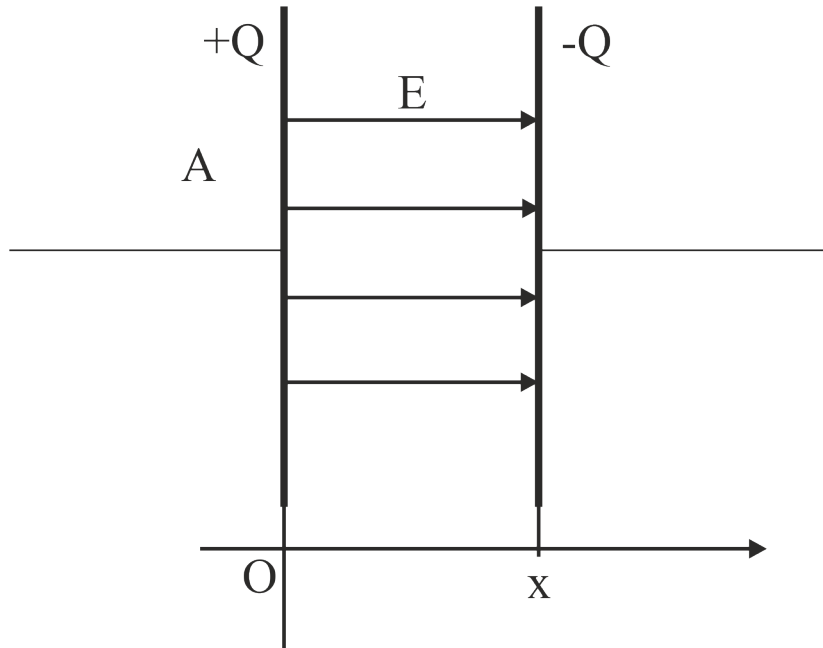


Figure 2.17: Determination of the electrostatic pressure

The following pieces of information are at disposal:

$$F = -\frac{dE_{pot}}{dx} \quad C(x) = \varepsilon_0 \varepsilon_r \frac{A}{x} \quad E_{pot} = \frac{1}{2} \frac{Q^2}{C} \quad (2.63)$$

The notations are as defined earlier. Let us substitute the formula of capacitance to the final formula:

$$E_{pot} = \frac{1}{2} \frac{Q^2}{C} = \frac{Q^2}{2} \frac{1}{C} = \frac{Q^2}{2} \frac{x}{\varepsilon_0 \varepsilon_r A} \quad (2.64)$$

Let us make the derivation. The absolute value of the force is as follows:

$$F = \frac{Q^2}{2A\varepsilon_0\varepsilon_r} \quad (2.65)$$

Thus between the plates of the capacitor a constant attractive force emerges. This is not surprising since opposite charges are facing each other at a little distance.

The following pieces of additional information are at disposal:

$$Q = CU \quad E = \frac{U}{d} \quad C = \varepsilon_0\varepsilon_r \frac{A}{d} \quad F = \frac{Q^2}{2A\varepsilon_0\varepsilon_r} \quad (2.66)$$

The notations are as defined earlier. The pressure is denoted p . Let us substitute to the final formula:

$$pA = F = \frac{Q^2}{2A\varepsilon_0\varepsilon_r} = \left(\varepsilon_0\varepsilon_r \frac{A}{d} \right)^2 U^2 \frac{1}{2A\varepsilon_0\varepsilon_r} = \frac{A}{2} \varepsilon_0\varepsilon_r \left(\frac{U}{d} \right)^2 = \frac{A}{2} \varepsilon_0\varepsilon_r E^2 \quad (2.67)$$

The pressure can readily be expressed:

$$p = \frac{1}{2} \varepsilon_0\varepsilon_r E^2 \quad \left[Pa = \frac{J}{m^3} \right] \quad (2.68)$$

This formula has shown up already in this chapter. The energy density and the pressure to the dielectric material between the plates are expressed by the same formula.

The critical electric field (E_{kr}) is the limit at which electric discharge occurs. The manufacturers of the capacitors carefully approach this limit by using tough materials. So the maximum pressure is determined approximately by the above formula at critical electric field intensity. For estimation purposes let us choose the following values: $\varepsilon_r = 3$ and $E_{cr} = 10^6$ V/m.

$$p_{\max} = \frac{1}{2} \varepsilon_0\varepsilon_r E_{cr}^2 = \frac{1}{2} 8.86 \cdot 10^{-12} \cdot 3 \cdot 10^{12} = 13.3 Pa \quad (2.69)$$

This pressure is an insignificant mechanical load on the dielectric material between the plates.

Chapter 3

Stationary electric current (direct current) - György Hárs

3.1 Definition of Ampere

Consider two pieces of metal electrodes on different potentials. The voltage between them is the difference of the potentials. Now connect the electrodes by means of a wire. The experiment proves that electric current flows on the wire as long as the voltage is sustained. The value of the current is the time derivative of the charge transferred. The unit of electric current is Ampere [A] which is a fundamental quantity in the SI system. Therefore the electric charge is a derived quantity and its unit is Ampere second [As] which can be called Coulomb.

$$I = \frac{dQ}{dt} \quad (3.1)$$

Currents in close proximity exert forces to each other. The definition of ampere is based on the force interaction between two parallel wires which carry the same current. It is worth mention here the important fact that parallel direction currents attract while the opposite direction currents repel each other. This is somewhat in contrary to the anticipation which might suggest otherwise.

It is also important to note that the direction of electric current is downhill the potential field. By definition the direction is from the plus to the minus electrode. And this is always true, no matter what kind of charge carrier is involved. If the charge carrier is negative (mostly electron) then the direction of mechanical flow is just opposite to the current direction.

Experiments show that the intensity of force (F) is proportional to the currents (I) and to the length (l) of the wire, while it is reversely proportional to the separation (r) of the parallel wires. To create equation from the proportionalities a coefficient is

introduced $(\mu_0/2\pi)$.

$$F = \frac{\mu_0}{2\pi} \frac{I^2}{r} l \quad (3.2)$$

The parameter μ_0 is a universal constant in nature and this is called the permeability of vacuum. The numerical value is $4\pi 10^{-7}$ Vs/Am.

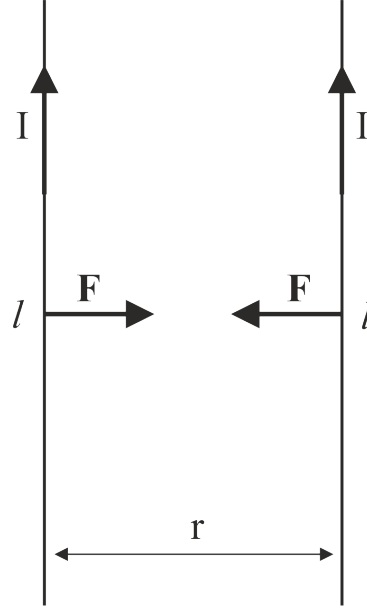


Figure 3.1: Attractive force between parallel currents

Based on the formula, the definition of Ampere is as follows: The values are 1A of two identical parallel currents if the attractive force is $2 \cdot 10^{-7}$ N between them provided both the length of the wire and the separation are one meter. The force to be measured is obviously very small, so much higher current and far smaller separation are used in the real measurements.

3.2 Current density (\mathbf{j})

The current density is a more essential physical quantity than the current itself. The current density is a vector. Its direction shows the local direction of the current. The flux of the current density on an open (S) surface results the actual current flowing through the rim of the surface. The measuring unit of the current density is A/m².

$$I = \int_S \mathbf{j} d\mathbf{A} \quad (3.3)$$

At homogeneous current density and plane surface the above integral can be replaced by the dot product of the current density and the corresponding area vector.

$$I = \mathbf{j} \cdot \mathbf{A} \quad (3.4)$$

If the current density vector and the area vector are parallel then the absolute value of the current density can be expressed from the equation:

$$j = \frac{I}{A} \quad (3.5)$$

3.3 Ohm's law

Experiments show that the current (I) is proportional to the voltage (U) applied. The coefficient between them is the conductance (G) of the conductor. The unit of the conductance is A/V called Siemens.

$$I = GU \quad (3.6)$$

The reciprocal value of the conductance is called the resistance (R). The unit of the resistance is V/A called Ohm (Ω).

$$R = \frac{1}{G} = \frac{U}{I} \quad (3.7)$$

The resistance of a cylindrical conductor is proportional to the length (l) and reversely proportional to the cross sectional area (A). The coefficient is characteristic to the material of the conductor which is called the resistivity (ρ). Its measuring unit is ohm meter (Vm/A).

$$R = \rho \frac{l}{A} \quad (3.8)$$

The reciprocal of the resistivity is called the conductivity (σ):

$$\sigma = \frac{1}{\rho} \quad (3.9)$$

Let us substitute to the Ohm's law:

$$U = RI = \rho \frac{l}{A} I \quad (3.10)$$

$$\frac{U}{l} = \rho \frac{I}{A} \quad (3.11)$$

The left hand side is the electric field (E) in the conductor while the right hand side is the current density (j) .

$$E = \rho j \quad (3.12)$$

This equation is the differential ohm's law. By means of conductivity the formula is as follows:

$$j = \sigma E \quad (3.13)$$

3.4 Joule's law

The power dissipated by the conductor is the time derivative of the work done by the electric field. The measuring unit is Watt ($\text{J/s} = \text{W}$).

$$P = \frac{dW}{dt} = U \frac{dQ}{dt} = UI \quad (3.14)$$

This formula is the Joule's law. Combining it with the Ohm's law the following formulas can be concluded.

$$P = UI = RI^2 = \frac{U^2}{R} \quad (3.15)$$

Let us use the formulas of U and I and substitute them to the above equation.

$$P = UI = (El)(jA) = Ej(Al) \quad (3.16)$$

The Al product is the volume of the conductor. After dividing with the volume, power density (p) can be expressed: The measuring unit is watt per cubic meter (W/m^3).

$$p = Ej \quad (3.17)$$

This formula is the differential Joule's law. Involving the differential Ohm's law two more expressions can be found:

$$p = \sigma E^2 \quad p = \rho j^2 \quad (3.18)$$

3.5 Microphysical interpretation

The charge carriers collide frequently with the ion lattice in the conductive material. Between collisions they are accelerated by the electric field. So the motion consists of short acceleration periods and sudden stops. The resulting motion can be characterized

by the average speed which is called the drift velocity (v_{drift}). Surprisingly this value is very small, roughly one meter per hour. Experiments show that the drift velocity is proportional to the electric field affecting the conductor. The coefficient is called the mobility (μ).

$$v_{drift} = \mu E \quad (3.19)$$

Consider a piece of conducting material with cross sectional area A . The material contains charge carriers with the density n and with charge q . The infinitesimal amount of charge transferred by the material in infinitesimal time period is as follows:

$$dQ = v_{drift} dt \cdot A \cdot nq \quad (3.20)$$

The current can be expressed:

$$\frac{dQ}{dt} = I = v_{drift} \cdot A \cdot nq \quad (3.21)$$

The current density can also be calculated:

$$j = v_{drift} \cdot nq \quad (3.22)$$

Now we substitute the mobility:

$$j = \mu nq \cdot E \quad (3.23)$$

Compare this result with the differential Ohm's law. This provides a microphysical substantiation to the conductivity, which was introduced earlier as a phenomenological material parameter.

$$\sigma = \mu nq \quad (3.24)$$

Accordingly the conductivity of some material depends on two major factors such as the mobility and the density of the charge carriers. The individual conductivity of all kinds of charge carriers are summarized provided several types of charge carriers are involved in the current.

If the temperature of the material is increased the conductivity can either increase or decrease. Increase in the conductivity occurred if the generation of the charge carriers is the dominant effect (mostly semiconductors). The conductivity decreases once the reduction of the mobility is the dominant effect (mostly metals).

Chapter 4

Magnetic phenomena in space - György Hárs

4.1 The vector of magnetic induction (B)

In chapter 3 the definition of ampere is based on the attractive force between two identical parallel currents. In this chapter the current values can be different. One of the currents is considered the source current (I) while the other one is the test current (i). Accordingly the intensity of force (F) is proportional to the currents (I, i) and to the length (l) of the wire, while it is reversely proportional to the separation (r) of the parallel wires. To create equation from the proportionalities a coefficient is introduced ($\mu_0/2\pi$).

$$F = \frac{\mu_0}{2\pi} \frac{Ii}{r} l \quad (4.1)$$

The parameter μ_0 is a universal constant in nature and this is called the permeability of vacuum. The numerical value is $4\pi 10^{-7}$ Vs/Am.

Experiments showed that the test current (i) and the length (l) of the test wire are proportional to the force. Accordingly the force can be written as follows:

$$F = \frac{\mu_0}{2\pi} \frac{Ii}{r} l = Bil \quad (4.2)$$

Here B is a coefficient which is determined solely by the source current (I), and somehow characteristic to the magnetization level of the space (magnetic induction field) generated by the source current.

$$\frac{\mu_0}{2\pi} \frac{I}{r} = B \quad \left[\frac{Vs}{m^2} = Tesla \right] \quad (4.3)$$

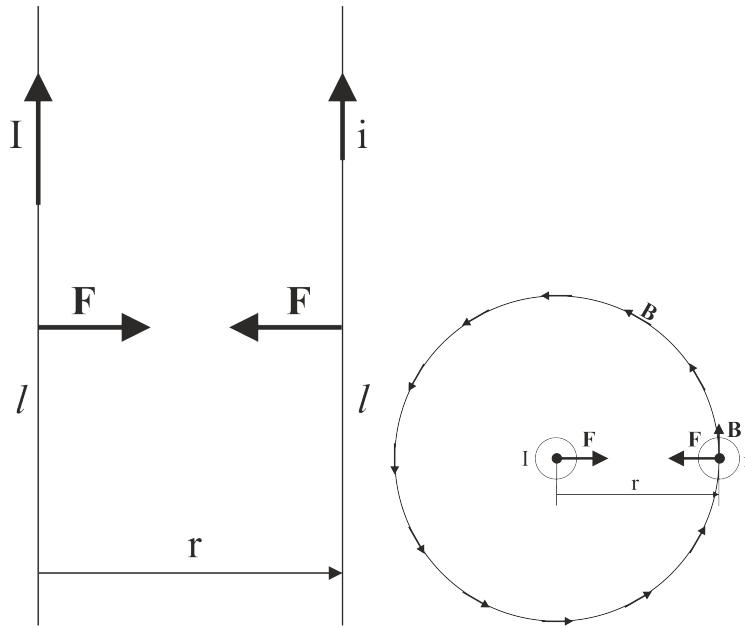


Figure 4.1: Parallel currents: side view(left) and upper view (right)

Up to this point this almost looks like if B were a scalar. This is not the case. The B will be in fact the vector of the magnetic induction (\mathbf{B}) with the definition below:

Let us study the upper view of the currents (Fig 4.2). Due to the cylindrical symmetry of the infinite straight current the magnetic induction lines are supposed to be circles around the current. Let us attribute right hand screw turning direction to the lines relative to the current direction. Accordingly when current flows out of the sheet the magnetic induction lines go around current counter clockwise (CCW).

4.2 The Lorentz force

On the other hand the attractive force vector points toward the source current. This direction is perpendicular both to the direction of \mathbf{B} vector and the test current. This relation implies the application of vector product as a mathematical means.

$$\mathbf{F} = i\mathbf{l} \times \mathbf{B} \quad (4.4)$$

The above formula accurately describes both the direction and the intensity of the force. The direction of the current is turned to the direction of the magnetic induction and the right hand screw turning will determine the direction of the force.

The test wire is not necessarily straight, this can be any curve. Very small (infinitesimal) section of a curve can be considered straight, so the above formula is valid for the

infinitesimal contribution to the force.

$$d\mathbf{F} = i d\mathbf{l} \times \mathbf{B} \quad (4.5)$$

The total force results as the curve integral of the contributions.

$$\mathbf{F} = \int_g i d\mathbf{l} \times \mathbf{B} \quad (4.6)$$

The validity of the earlier formula can be extended to point charges traveling in the space. If a point charge moves this is equivalent with a certain current. This relation is summarized below:

$$i d\mathbf{l} = \mathbf{j} A \cdot d\mathbf{l} = \mathbf{v} \rho \cdot A d\mathbf{l} = \mathbf{v} \cdot \rho A d\mathbf{l} = \mathbf{v} dQ \quad (4.7)$$

Here we used the expression of current density by means of the charge density and the velocity: $\mathbf{j} = \rho \mathbf{v}$

The above expression can be substituted to expression of force:

$$d\mathbf{F} = dQ(\mathbf{v} \times \mathbf{B}) \quad (4.8)$$

In case of point charge there is a definite amount of charge and so the force is a definite vector too.

$$\mathbf{F} = Q(\mathbf{v} \times \mathbf{B}) \quad (4.9)$$

This formula describes the Lorentz force. Accordingly the magnetic field may affect a charged particle only when it moves. Standstill particle does not “feel” the magnetic field. If the above formula is divided with the charge the Lorentz electric field (\mathbf{E}_L) is the result.

$$\mathbf{E}_L = \mathbf{v} \times \mathbf{B} \quad (4.10)$$

This quantity will be used extensively in connection with the motion related electromagnetic induction phenomena.

4.2.1 Cyclotron frequency

Consider homogeneous magnetic field ($B = 0.1 \text{ Tesla}$). Inject a proton ($m_p = 1.67 \cdot 10^{-27} \text{ kg}$, $q_p = 1.6 \cdot 10^{-19} \text{ As}$) normal to the magnetic field with initial kinetic energy ($U_0 = 1 \text{ keV}$). Determine how the particle moves in the field.

The Lorentz force is always normal to the velocity therefore the speed (and so the kinetic energy) of the particle is constant. The Lorentz force generates only centripetal

acceleration, and this way the particle goes around a circular trajectory. The parameters of the motion can be determined by means of the equation of motion:

$$qvB = m \frac{v^2}{r} \quad (4.11)$$

The equation above is written in the radial direction of the circle. The left hand side is the absolute value of the Lorentz force while the right hand side is the mass multiplied with the centripetal acceleration. After some ordering:

$$\frac{qB}{m} = \frac{v}{r} = \omega_{cyclotron} = \frac{1.6 \cdot 10^{-19} \cdot 0.1}{1.67 \cdot 10^{-27}} = 9.58 \cdot 10^6 \frac{rad}{s} = 1.53 MHz \quad (4.12)$$

The velocity can be calculated from the initial kinetic energy:

$$\frac{1}{2}mv^2 = qU_0 \quad (4.13)$$

$$v = \sqrt{\frac{2qU_0}{m}} = \sqrt{\frac{2 \cdot 1.6 \cdot 10^{-19} \cdot 10^3}{1.67 \cdot 10^{-27}}} 4.38 \cdot 10^5 \frac{m}{s} \quad (4.14)$$

The radius of the circulation can be expressed:

$$r = \frac{mv}{qB} = \frac{v}{\omega_c} = \frac{4.38 \cdot 10^5}{9.58 \cdot 10^6} = 4.57 \cdot 10^{-2} m = 4.57 cm \quad (4.15)$$

Note the cyclotron frequency is independent of the energy of the particle. This feature made possible to construct the first particle accelerator (1932 Ernest Lawrence). The charged particles are forced to circulate by means of homogeneous magnetic field. They are accelerated with a high frequency electric field which is in resonance with the cyclotron frequency. As the energy of the particles grew the radius of the circulation increased but the cyclotron frequency did not change so the resonance stayed. The particles could be accelerated as long as the classical approach is worked. At higher energies the relativistic description is necessary.

If the injection of the particle is not fully perpendicular to the \mathbf{B} vectors then the initial velocity should be decomposed to parallel and normal components. The parallel component is unaffected by the magnetic field while the normal component generates uniform circulation with the cyclotron frequency. Ultimately the trajectory of the particle is twisted around the magnetic induction lines. This feature is used extensively in plasma generation techniques when additional external magnetic field is used to increase the efficiency of the ionization by increasing the path length of the charged particle. Longer the path length higher is the probability of the collisions thus the ionization.

4.2.2 The Hall effect

The effect was discovered by Edwin Hall in 1879. Consider a layer of a conducting material in the form of a stripe. The direct current (I) flows parallel with the longer dimension. Homogeneous magnetic field (B) crosses the material normal to the surface. The Hall voltage is measured between the two sides of the stripe.

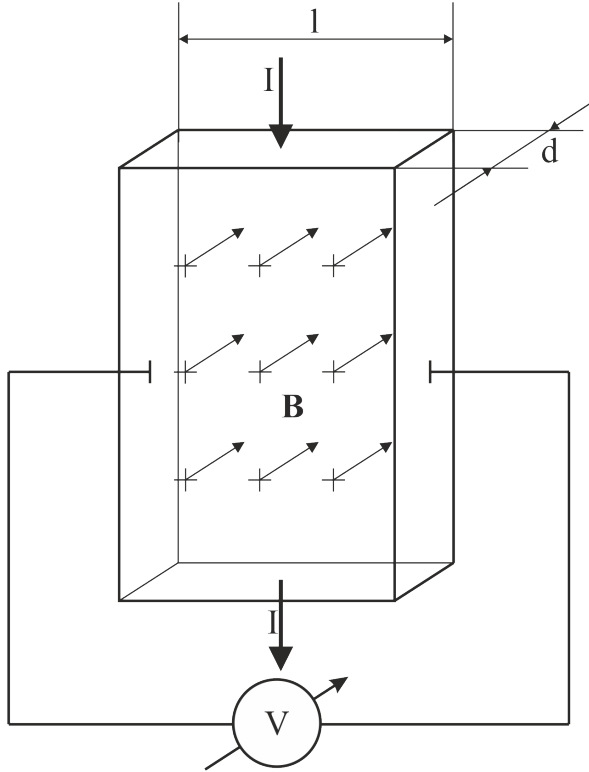


Figure 4.2: Hall effect

The following pieces of information are at disposal:

$$I = j \cdot l \cdot d \quad [A] \quad j = v_{drift} \cdot nq \quad \left[\frac{A}{m^2} \right] \quad (4.16)$$

Here l and d are the width and the thickness of the stripe respectively.

$$I = v_{drift} \cdot nq \cdot l \cdot d \quad (4.17)$$

The drift velocity can be expressed:

$$v_{drift} = \frac{I}{nq \cdot l \cdot d} \quad (4.18)$$

The absolute value of the Lorentz electric field is merely the product of the factors due to the perpendicular arrangement.

$$E_L = v_{drift} B \quad (4.19)$$

The generated Hall voltage is the product of the width (l) and the Lorentz field intensity. Integration can be omitted because the field is homogeneous.

$$U_L = v_{drift} B \cdot l = \frac{BI}{nq \cdot d} = \frac{1}{nq} \frac{BI}{d} = R_H \frac{BI}{d} \quad (4.20)$$

Here R_H is the Hall coefficient as follows:

$$R_H = \frac{1}{nq} \left[\frac{m^3}{As} \right] \quad (4.21)$$

The formula shows that the polarity of the Hall coefficient depends on the charge carrier polarity. In all other experiments electrons travel from left to right makes the same effect when positive particles travel from the opposite direction. So one never knows just from the current, which is the case in fact. This is the only experiment in which the polarity of the charge carrier makes a qualitative difference.

In modern electronic devices Hall detector is used mostly for measuring magnetic field. It is also used as commutators in electric motors and in the ignition system of the cars.

4.3 Magnetic dipole

Consider a circular loop current (I) with given radius (r) in the x, y plane of the Cartesian coordinate system. The loop current is surrounded by homogeneous magnetic (\mathbf{B}) field in arbitrary direction. The infinitesimal force vector affecting an infinitesimal section of the loop is as follows:

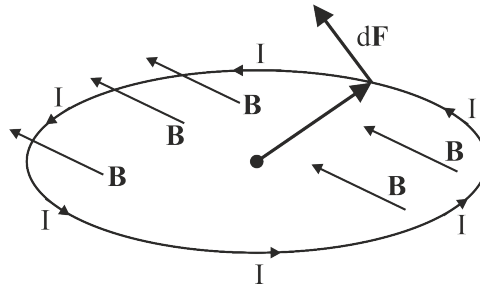


Figure 4.3: Determination of the torque affecting a current loop

$$d\mathbf{F} = Id\mathbf{l} \times \mathbf{B} \quad (4.22)$$

The infinitesimal contribution of torque is based on the definition:

$$d\mathbf{M} = \mathbf{r} \times d\mathbf{F} = \mathbf{r} \times (Id\mathbf{l} \times \mathbf{B}) \quad (4.23)$$

Now we use the formula for triple product: $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b})$ Accordingly:

$$d\mathbf{M} = Id\mathbf{r}(\mathbf{B} \cdot \mathbf{r}) - \mathbf{B}(\mathbf{r} \cdot d\mathbf{r}) \quad (4.24)$$

The second term is zero because the \mathbf{r} and $d\mathbf{r}$ vectors are perpendicular. So ultimately the torque to be integrated is the following: $d\mathbf{M} = Id\mathbf{r}(\mathbf{B} \cdot \mathbf{r})$

$$\mathbf{M} = \int_0^{2\pi} d\mathbf{M} = I \int_0^{2\pi} (\mathbf{B} \cdot \mathbf{r}) d\mathbf{r} \quad (4.25)$$

Next the formulas to be substituted:

$$\mathbf{B} = B_x \mathbf{i} + B_y \mathbf{j} + B_z \mathbf{k} \quad (4.26)$$

$$\mathbf{r} = \mathbf{i} \cdot r \cos \phi + \mathbf{j} \cdot r \sin \phi + 0 \cdot \mathbf{k} \quad (4.27)$$

$$d\mathbf{r} = \frac{d\mathbf{r}}{d\phi} d\phi = r(-\mathbf{i} \cdot \sin \phi + \mathbf{j} \cdot \cos \phi) d\phi \quad (4.28)$$

The formula is simplified by the following rule: $\mathbf{i}\mathbf{j} = \mathbf{j}\mathbf{k} = \mathbf{k}\mathbf{i} = 0$ and $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = 1$

$$\mathbf{M} = Ir^2 \int_0^{2\pi} (B_x \cos \phi + B_y \sin \phi)(-\mathbf{i} \cdot \sin \phi + \mathbf{j} \cdot \cos \phi) d\phi \quad (4.29)$$

The formula is further simplified by the orthogonal sine and cosine functions. The only remaining terms are either pure sine or pure cosine functions.

$$\mathbf{M} = Ir^2 \left(\int_0^{2\pi} (B_x \cos^2 \phi \cdot d\phi) \mathbf{j} - \int_0^{2\pi} (B_y \sin^2 \phi \cdot d\phi) \mathbf{i} \right) \quad (4.30)$$

The values of the integral can be calculated:

$$\int_0^{2\pi} \cos^2 d\phi = \int_0^{2\pi} \sin^2 d\phi = \pi \quad (4.31)$$

$$\mathbf{M} = Ir^2 (B_x \pi \cdot \mathbf{j} - B_y \pi \cdot \mathbf{i}) = Ir^2 \pi (-B_y \cdot \mathbf{i} + B_x \cdot \mathbf{j}) \quad (4.32)$$

Here one can discover the area of the circle. $A = r^2 \pi$

Finally the formula of torque emerges.

$$\mathbf{M} = IA \cdot (-B_y \cdot \mathbf{i} + B_x \cdot \mathbf{j}) \quad (4.33)$$

This formula is not easy to handle let alone to remember that. If one attributes vector character to the area as already had been, the above formula can be interpreted much more elegant way.

$$\mathbf{M} = I \mathbf{A} \times \mathbf{B} \quad (4.34)$$

Let us check this formula.

$$\mathbf{M} = I \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 0 & A \\ B_x & B_y & B_z \end{vmatrix} = IA \cdot (-B_y \cdot \mathbf{i} + B_x \cdot \mathbf{j}) \quad (4.35)$$

This is a perfect match.

Due to historical reasons the formula of torque is modified with the permeability of vacuum. :

$$\mathbf{M} = \mu_0 I \mathbf{A} \times \frac{\mathbf{B}}{\mu_0} \quad (4.36)$$

The first factor in the cross product is the magnetic dipole moment (\mathbf{m}) of the current loop.

$$\mathbf{m} = \mu_0 I \mathbf{A} \quad [Vsm] \quad (4.37)$$

The second factor is denoted \mathbf{H} called magnetic field measured in A/m unit. Its physical meaning will be explained later in the next chapter.

$$\mathbf{M} = \mathbf{m} \times \mathbf{H} \quad (4.38)$$

The magnetic dipole moment is turned into the direction of the external magnetic field spontaneously and stays there. Having reached this position, the least amount of potential energy is stored in the magnetic dipole. Obviously the most amount of potential

energy stored is just in the opposite position. Let us find out the work needed to turn the dipole from the deepest position to the highest energy.

$$W = \int_0^\pi M d\phi = \int_0^\pi mH \sin \phi \cdot d\phi = mH \int_0^\pi \sin \phi \cdot d\phi = -mH [\cos \phi]_0^\pi = 2mH \quad (4.39)$$

According to this result the potential energy of the magnetic dipole is as follows:

$$E_{pot} = -\mathbf{m} \cdot \mathbf{H} \quad (4.40)$$

This formula provides the deepest energy at parallel spontaneous position and the highest at anti-parallel position. The zero potential energy is at ninety degrees. The difference between the highest and lowest is just the work needed to turn it around.

4.4 Earth as a magnetic dipole

Magnetic dipole moment of a solenoid is directed according to the right hand screw rule. This means that rotating parallel with the circulation of the current, the progress of the right hand screw defines the direction of the magnetic dipole moment. Let us suspend the solenoid in its center of gravity on a thin thread which provides free turning in the horizontal plane. The solenoid slowly turns parallel to the Earth's magnetic field such a way that the magnetic dipole moment points to the geographic North Pole. If a permanent magnet rod is suspended in the same way as the solenoid, this will also turn parallel to the Earth's magnetic field.

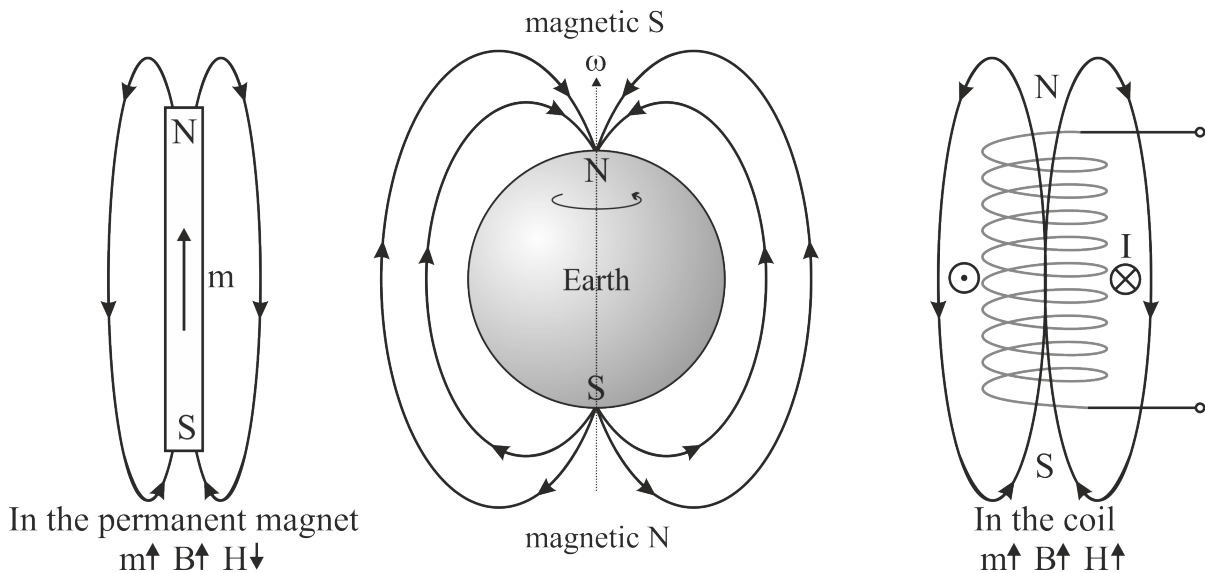


Figure 4.4: The geographic North Pole of the Earth is in fact a magnetic South Pole

In case of an ordinary dipoles such as a solenoid or a bar magnet the magnetic dipole moment is directed from the south end to the north end. The magnetic induction lines are virtually exiting from the north end and entering into the south end.

Planet Earth is an exceptional magnetic dipole because the geographic North Pole is in fact a magnetic South Pole. This weird-looking switch is required to dissolve the contradiction of naming the poles of an ordinary dipole. That end of an ordinary dipole is called north end which is closer to the north geographic pole of the Earth. However the same kinds of poles repel each other so the mentioned switch clears the situation.

4.5 Biot-Savart law

The magnetic field of an infinitesimal current element is described by the Biot-Savart law.

$$d\mathbf{H} = \frac{I}{4\pi} \frac{d\mathbf{l} \times \mathbf{r}_0}{r^2} \quad (4.41)$$

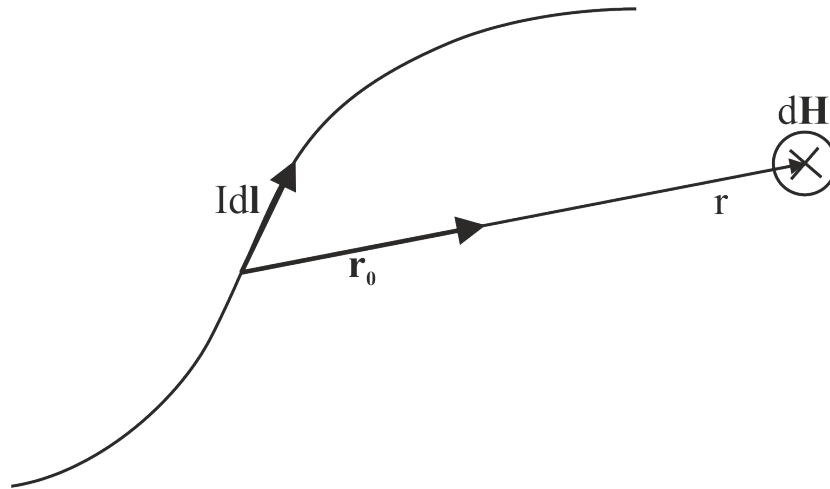


Figure 4.5: Biot-Savart law

The infinitesimal current element is surrounded by circular magnetic field. The rotation of the magnetic field is in accordance with the right hand screw rule. In the equatorial plane the magnetic field diminishes with the negative second power (just like Coulomb's law). Below and above the equatorial plane the magnetic field diminishes with the increasing angle and vanishes on the line of the current element. The infinitesimal magnetic contributions can be superimposed and the overall magnetic effect of any extended current can be calculated by integration. The Biot-Savart law is used typically for currents in a thin wire where the integration by line provides a fair result.

4.5.1 Magnetic field of the straight current

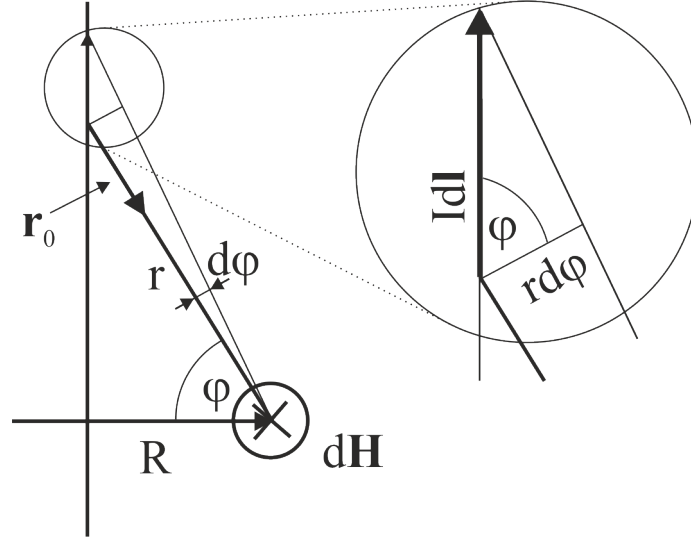


Figure 4.6: Biot-Savart law for a straight wire

The following pieces of information are at disposal:

$$d\mathbf{H} = \frac{I}{4\pi} \frac{d\mathbf{l} \times \mathbf{r}_0}{r^2} \quad |d\mathbf{l}| = \frac{rd\phi}{\cos \phi} \quad r = \frac{R}{\cos \phi} \quad (4.42)$$

All contributions of the magnetic field point to the same direction therefore the integration of the absolute value is satisfactory.

$$|d\mathbf{H}| = \frac{I}{4\pi} \frac{|d\mathbf{l}| \cdot |\mathbf{r}_0| \sin(\phi + 90^\circ)}{r^2} = \frac{I}{4\pi} \frac{rd\phi \cos \phi}{\cos \phi} \frac{1}{r^2} = \frac{I}{4\pi} \frac{d\phi}{r} = \frac{I}{4\pi} \frac{\cos \phi}{R} d\phi \quad (4.43)$$

After some simplifications the infinitesimal contribution results:

$$dH = \frac{I}{4R\pi} \cos \phi \cdot d\phi \quad (4.44)$$

The integration will be carried out for symmetrical α half visual angle domain.

$$H = \int_{-\alpha}^{\alpha} \frac{I}{4R\pi} \cos \phi \cdot d\phi = \frac{I}{4R\pi} [\sin \phi]_{-\alpha}^{\alpha} = \frac{I}{2R\pi} \sin \alpha \quad (4.45)$$

The magnetic field of a finite current section under symmetrical α half visual angle domain can finally be expressed.

$$H = \frac{I}{2R\pi} \sin \alpha \quad (4.46)$$

If the current tends to the infinity then α tends to ninety degrees so $\sin\alpha$ equals unit. This way for infinite long filament the result is as follows:

$$H = \frac{I}{2R\pi} \quad (4.47)$$

4.5.2 Central magnetic field of the polygon and of the circle

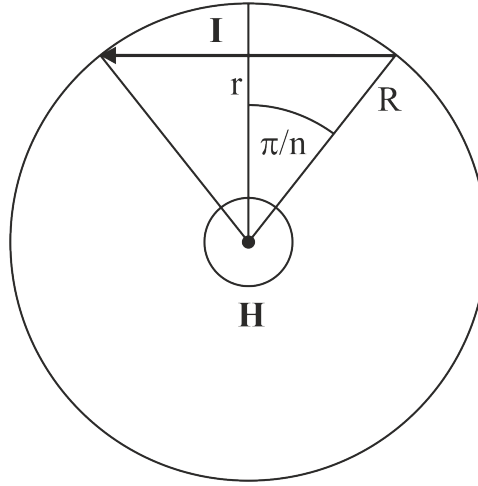


Figure 4.7: Regular polygon

Consider a part of a regular polygon with sides n which carries a current I . The containing circle has the radius R . The half central angle is π/n . The magnetic effect of one side is denoted H^* . For other notations see the figure above.

The following pieces of information are at disposal:

$$H^* = \frac{I}{2r\pi} \sin\left(\frac{\pi}{n}\right) \quad r = R \cos\left(\frac{\pi}{n}\right) \quad H_n = n \cdot H^* \quad (4.48)$$

After substitution:

$$H_n = n \frac{I}{2\pi} \frac{1}{R \cos\left(\frac{\pi}{n}\right)} \sin\left(\frac{\pi}{n}\right) = \frac{I}{2R} \left(\frac{n}{\pi}\right) \frac{\sin\left(\frac{\pi}{n}\right)}{\cos\left(\frac{\pi}{n}\right)} = \frac{I}{2R} \left(\frac{n}{\pi}\right) \operatorname{tg}\left(\frac{\pi}{n}\right) \quad (4.49)$$

So altogether the magnetic field in the center of the n sided regular polygon is as follows:

$$H_n = \frac{I}{2R} \left(\frac{n}{\pi}\right) \operatorname{tg}\left(\frac{\pi}{n}\right) \quad (4.50)$$

Please find some numerical values:

$$H_3 = \frac{I}{2R} \left(\frac{3}{\pi}\right) \operatorname{tg}60^\circ = 1.65 \frac{I}{2R} \quad (4.51)$$

$$H_4 = \frac{I}{2R} \left(\frac{4}{\pi} \right) tg90^0 = 1.27 \frac{I}{2R} \quad (4.52)$$

$$H_5 = \frac{I}{2R} \left(\frac{5}{\pi} \right) tg36^0 = 1.15 \frac{I}{2R} \quad (4.53)$$

$$H_6 = \frac{I}{2R} \left(\frac{6}{\pi} \right) tg30^0 = 1.10 \frac{I}{2R} \quad (4.54)$$

Let us make a transformation on the original result:

$$H_n = \frac{I}{2R} \left(\frac{n}{\pi} \right) tg\left(\frac{\pi}{n}\right) = \frac{I}{2R} \left(\frac{tg\left(\frac{\pi}{n}\right)}{\left(\frac{\pi}{n}\right)} \right) \quad (4.55)$$

If the number of sides tends to the infinity then the polygon will tend to the circle. The limit value of the big parenthesis is unit. Ultimately the magnetic field in the center of the circle is equal with the current over the diameter (worth remembering).

$$H_{circle} = H_{\infty} = \frac{I}{2R} \quad (4.56)$$

4.6 Ampere's law

The magnetic field of the infinite long filament was reached in this chapter above.

$$H = \frac{I}{2r\pi} \quad (4.57)$$

One can transform this formula in the following way:

$$2r\pi \cdot H(r) = I \quad (4.58)$$

If the circumference of a circle is multiplied with the actual magnetic field, the result will be independent of the radius and will be equal with the current. Since the magnetic field is constant on a radius one may write the above equation also by means of curve integral on the circle.

$$\oint_{circle} \mathbf{H}(\mathbf{r}) d\mathbf{r} = I \quad (4.59)$$

This equation prompts the following hypothesis. May be the integration path needs not to be a circle but this could be any closed loop around the current. This hypothesis is

proven and it is called the Ampere's law. The current may even be the sum of several currents. The direction of integration determines the direction of positive currents based on the right hand screw rule.

$$\oint_{loop} \mathbf{H}(\mathbf{r}) d\mathbf{r} = \sum I \quad (4.60)$$

The proof the Ampere's law next:

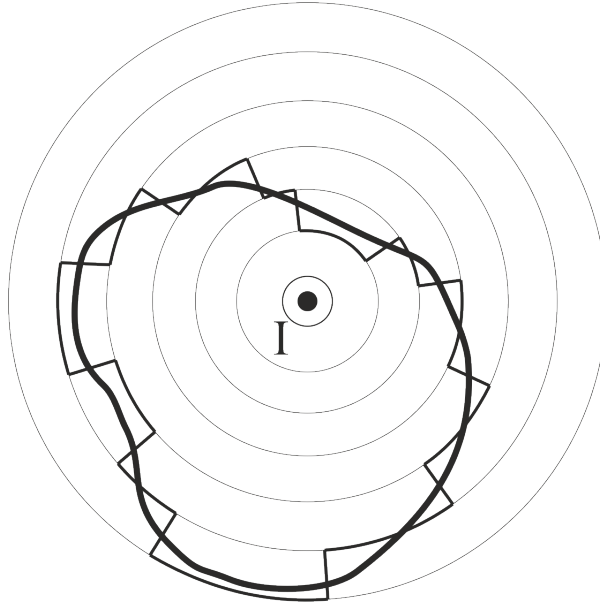


Figure 4.8: Ampere's law proof

The figure shows an infinite straight current normal to the sheet of paper pointing up just in front of our eyes. An arbitrary closed loop surrounds the current which is the path of the line integral. The current is also surrounded by several concentric circles in equidistant steps. The path of the integration can be approximated by very small (infinitesimal) sections which are either in tangential or in radial positions relative to the concentric circles. This way the integration can be carried out by moving on any of the circles or by moving radial direction. No contribution is generated in radial direction since the dot product vanishes at perpendicular position. On the circles however the contribution is the product of the magnetic field and the length of the arc. The corresponding central angle is donated φ .

$$\oint_{loop} \mathbf{H}(\mathbf{r}) d\mathbf{r} = (\phi_1 r_1) \cdot \frac{I}{2r_1\pi} + (\phi_2 r_2) \cdot \frac{I}{2r_2\pi} + (\phi_3 r_3) \cdot \frac{I}{2r_3\pi} +(\phi_n r_n) \cdot \frac{I}{2r_n\pi} \quad (4.61)$$

The radii all cancel out.

$$\oint_{loop} \mathbf{H}(\mathbf{r}) d\mathbf{r} = \frac{I}{2\pi} (\phi_1 + \phi_2 + \phi_3 + \dots + r_n) = I \quad (4.62)$$

The sum of the central angles stacks up to 2π due to the closed loop. So altogether the angles cancel out. Ultimately the statement to be proven is the result. Q.E.D.

Application of Ampere's law for solving problems requires similar considerations to that of the Gauss's law. Ampere's law is an integral law. If one wants to use it for finding out local magnetic field, the symmetries or regularities of the magnetic field must be known prior to the application. If so, one has to choose the path of the integration in which the magnetic field is constant thus the integral converts to a simple product. The actual magnetic field results after dividing with the length of the integration path.

4.6.1 Thick rod with uniform current density

Consider a thick metal rod with radius R , which conducts current with uniform current density denoted j . Find the intensity of the magnetic field as the function of radius.

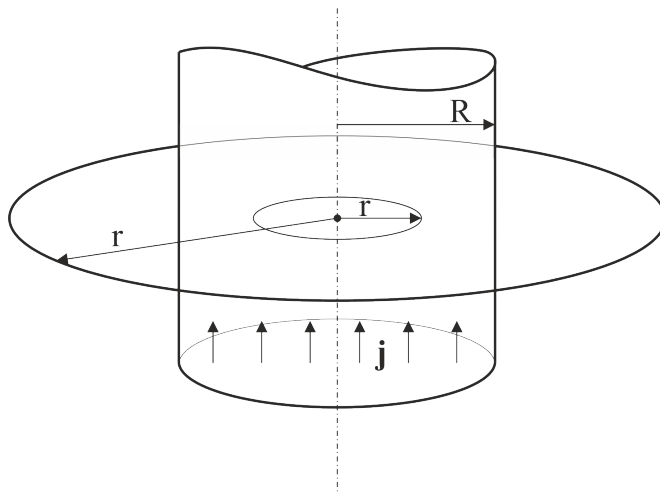


Figure 4.9: Current in the thick rod

The solution uses Ampere's law. A circle is inflated in radius from zero to the infinity.

$$\oint_{loop} \mathbf{H}(\mathbf{r}) d\mathbf{r} = \sum I \quad (4.63)$$

In the rod:	Out of the rod
$2r\pi \cdot H = r^2\pi \cdot j$	$2r\pi \cdot H = R^2\pi \cdot j$
$H = \frac{j}{2}r$	$H = \frac{j}{2}\frac{R^2}{r}$

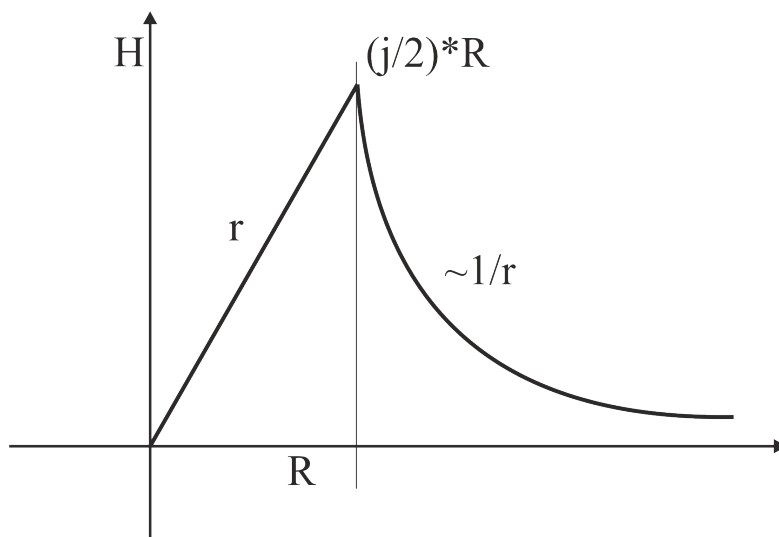


Figure 4.10: $H(r)$ function

Inside there is a linear slope of the magnetic field intensity. Outside, the intensity decays like a hyperbola. The function is continuous on the surface.

4.6.2 Solenoid

This is a straight rod coil. The physical model of solenoid requires the length to be roughly ten times longer than the diameter.

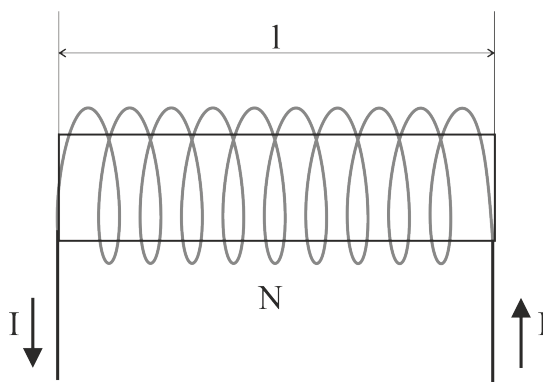


Figure 4.11: The solenoid coil

Let us use Ampere's law. The magnetic field is homogeneous in the cavity of the solenoid. A closed path is chosen which is parallel with the coil and located in the cavity on one side. The front side of the path is outside the coil where there is no magnetic field. The remaining two little sides of the rectangle are normal to the magnetic field thus can be ignored. The length is denoted l and the number of turns is N . So ultimately the Ampere's law emerges in a simple form:

$$\oint_{loop} \mathbf{H}(\mathbf{r}) d\mathbf{r} = \sum I \quad (4.64)$$

$$Hl = NI \quad (4.65)$$

$$H = \frac{NI}{l} \quad (4.66)$$

4.6.3 Toroidal coil

This is a doughnut shape coil. The physical model requires the circumference to be roughly ten times longer than the diameter.

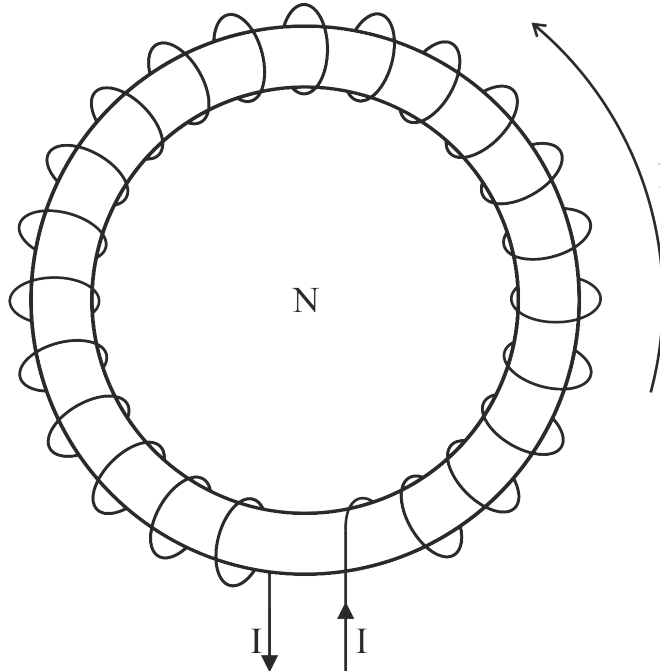


Figure 4.12: The toroidal coil

Let us use Ampere's law. The magnetic field is homogeneous in the cavity of the coil. A closed circular path is chosen which is running in the central of the cavity. The circumference is denoted l and the number of turns is N . So ultimately the Ampere's law emerges in a simple form:

$$\oint_{loop} \mathbf{H}(\mathbf{r}) d\mathbf{r} = \sum I \quad (4.67)$$

$$Hl = NI \quad (4.68)$$

$$H = \frac{NI}{l} \quad (4.69)$$

4.7 Magnetic flux

The general concept of flux has been treated earlier. This is a scalar value surface integral of some vector field. The vector field in present case is the field of magnetic induction $\mathbf{B}(\mathbf{r})$. The magnetic flux as follows:

$$\Phi_m = \int_g \mathbf{B} d\mathbf{A} \quad [Vs] \quad (4.70)$$

In case of solenoid and toroid the formula of the magnetic turn flux is as follows:

$$B = \frac{\mu_0 NI}{l} \quad \Phi_{turn} = BA = \frac{\mu_0 NIA}{l} \quad (4.71)$$

Later coil flux will also be used in conjunction with the induced voltage of the coil. :

$$\Phi_{coil} = \frac{\mu_0 N^2 IA}{l} \quad (4.72)$$

Chapter 5

Magnetic field and the materials - György Hárs

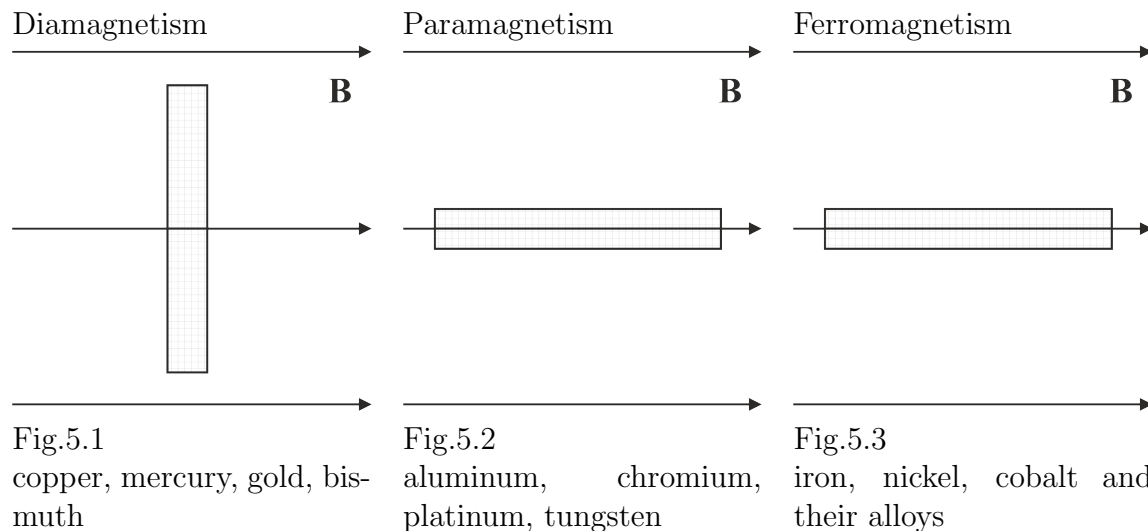
In the former chapter magnetic phenomena without any magnetic material in the surrounding space have been treated. In technology however, the magnetic field is used in conjunction with magnetic materials, which enhance the intensity of the forces and the related interactions.

5.1 Three basic types of magnetic behavior

Non-physicist public opinion divides the materials such as magnetic and non-magnetic materials. The former one is mostly iron and some alloys which are attracted by magnet and the latter ones are the rest of the materials which are seemingly unaffected by magnet. In reality however, all materials are affected by the magnetic field though the intensity of the attraction varies several orders of magnitude. The commonly mentioned “magnetic materials” are in fact the ferromagnetic materials in technical terms. The nonmagnetic materials can be classified to two distinct groups such as paramagnetic and diamagnetic materials in which the intensity of the interaction is so low that it is simply overlooked by the easy observer.

The following experiment makes it possible to distinguish between the sorts of the magnetic behavior: Make little samples of the materials to be tested. The sample geometry is roughly fifty millimeter long and five millimeter in diameter cylindrical rod. In the symmetry axis there is an indentation normal to the rotational axis which contains a needle on which the rod can be rotated freely. The depth of the indentation is roughly eighty percent of the rod diameter. The rotatable sample is placed into the air-gap of the unexcited toroidal electromagnet which can create high intensity homogeneous magnetic field. Now switch the electromagnet which creates a magnetic induction in the order of magnitude 0.01 Tesla at least. The samples of different materials will behave as shown

in the figures below:



Some samples orient themselves diagonally (perpendicularly) to the direction of the magnetic field. These materials are called diamagnetic materials (Fig 1). Some other materials orient themselves parallel to the magnetic field. These are the paramagnetic (Fig 2) and ferromagnetic (Fig 3) materials. The intensity of the interaction can be estimated based on the dynamics of the turning. The most sluggish turning happened at the diamagnetic materials. The paramagnetic material turned somewhat more agile but still slow. The turning reaction of ferromagnetic material was instantaneous relative to the others. The estimated intensities of the torques are roughly one, ten and several millions respectively.

So far the experimental distinction has been carried out. The microphysical interpretation of the experimental results should provide the understanding of the phenomena. The roots of the interpretation come from the atomic structure of the material. Some materials contain atoms without any magnetic dipole moment. Upon placing these materials into the magnetic field the atoms become weak dipoles. Due to quantum mechanical reasons the direction of the dipole moment will be just opposite of the external magnetic field. This way the sample becomes a dipole of opposite direction to the external magnetic field. Now the external field wants to turn the dipole parallel with itself. As the turning goes on the atomic dipoles also turn their orientation again against the external field. So the only position where the sample gets rest is the perpendicular position where virtually no torque affects the sample. This behavior is characteristic of the diamagnetic materials.

Atomic dipoles are originally located in the paramagnetic materials. Once external magnetic field emerges, the atomic dipoles in the material orient themselves into the

direction of the field, thus the sample becomes a magnetic dipole. In the former chapter the fact has been presented that the magnetic dipoles turn to the parallel direction to the external field. This way the sample is positioned as shown in the figure two. The diamagnetic effect also shows up in paramagnetic materials but it is overcompensated by the paramagnetic effect.

From technical point of view the most important type, the ferromagnetic material is treated finally. There are atomic dipoles in the material similar to the paramagnetic materials, but these atomic dipoles interact with each other in contrast to the simple paramagnetic case where the dipoles are affected by the external field alone. Due to the interaction, the dipoles orient themselves parallel with each other and create the magnetic domains. The size of such domains is roughly in the order of micrometers, which in terms of atomic dimensions is large, though in terms of macroscopic dimensions is still rather small. Such material is magnetically neutral since the domains are oriented randomly, this way the effects of domains average out to zero. When the external magnetic field is switched on the domains get oriented, and very high magnetic field is generated. The magnetization of the material is proportional to the relatively low external magnetic fields. At high external fields however the magnetization gets saturated since the domains have all been oriented. This phenomenon called the hysteresis. Another interesting fact is related to the Curie temperature. Above this temperature the material loses the ferromagnetic behavior and reverts to be paramagnetic. In the case of iron the Curie temperature is 770 degrees Celsius. The relatively high temperature breaks the bonds of the interaction between the dipoles, and domains are disintegrated.

Later in this work the discussion of magnetic phenomena is limited to ferromagnetic materials in the linear magnetization range when the magnetization and the external field are proportional. Hysteresis phenomena will not be treated in detail.

5.2 Solenoid coil with iron core

Take a solenoid coil with empty cavity. Switch on the DC current and feel how strong the magnetic force is by approaching it with an iron screwdriver. Now place the iron core into the cavity and check the force again. The experiment confirms that the force is much stronger than previously. The conclusion can be drawn readily that the presence of iron core increased the intensity of the force, in contrast to the electrostatic phenomena where the presence of dielectric material diminished the intensity of the force. This anti-symmetry is rooted in the fact that the parallel currents and the opposite charges attract each other.

A simple model will be discussed here suggested by Ampere:

The atomic dipole is equivalent with a tiny loop current. So the magnetic material is assumed to be filled with such loop currents in random orientation, so their total magnetic moment is zero. By switching the external DC field on, they get oriented and the loop

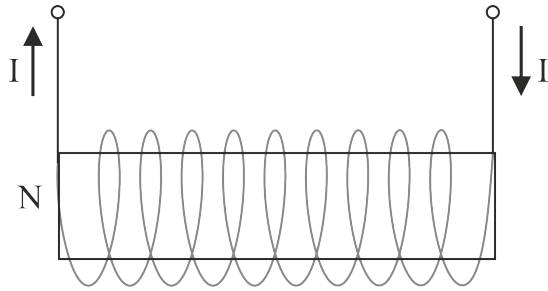


Fig.5.4
Solenoid from side view

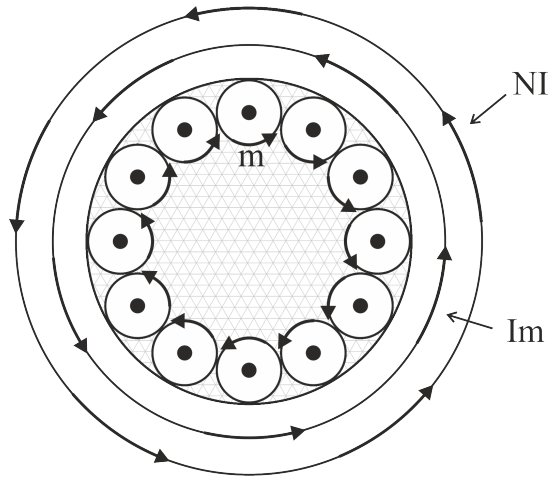


Fig.5.5
Solenoid from top view

currents will all circulate in the same direction. Let us consider a point within the iron core. Due to symmetry reasons same amount of current goes through this point from right to left and opposite. Inside the volume of the core the effects of loop currents neutralize each other. The situation is very much different on the surface of the core where the direction of the loop currents is parallel with the coil current. Because of this, the core will behave as it was a coil in which the so called magnetizing current (I_m) flows. The direction of the coil current and the direction of magnetizing current are the same.

The magnetic field of both the coil current and that of magnetizing current are as follows:

$$H = \frac{NI}{l} \quad H_m = \frac{I_m}{l} \quad (5.1)$$

Here N is the number of turns, I and I_m are the coil current and the magnetizing current finally the l is the length of the solenoid.

Due to the same direction the total magnetic field is the sum of these:

$$H_{tot} = H + H_m \quad (5.2)$$

Let us multiply the above equation with μ_0 .

$$\mu_0 H_{tot} = \mu_0 H + \mu_0 H_m \quad (5.3)$$

The left hand side of the equation is the field of the magnetic induction B . The second term on the right hand side is magnetization M , accordingly this can be written:

$$B = \mu_0 H + M \quad (5.4)$$

The deduction of this formula was made in a special geometry for the sake of simplicity. However the result is true in full generality for vectors as well.

$$\mathbf{B} = \mu_0 \mathbf{H} + \mathbf{M} \quad \left[\frac{Vs}{m^2} \right] \quad (5.5)$$

The meaning of the above formula can be summarized as follows:

The magnetic field (\mathbf{H}) is generated by the coil current alone, while the vector of magnetization (\mathbf{M}) is generated solely by the magnetizing current. The vector of magnetic induction (\mathbf{B}) contains the effects of both the coil current and the magnetizing current. The emerging torques and forces are determined by the \mathbf{B} field.

The experiment showed that the created magnetization (\mathbf{M}) is proportional to the external field (\mathbf{H}) provided no saturation happens. Proportionality can be transformed to equation by introducing a coefficient which is call the magnetic susceptibility (χ_m).

$$\mathbf{M} = \mu_0 \chi_m \mathbf{H} \quad (5.6)$$

Let us substitute this into the former one:

$$\mathbf{B} = \mu_0 \mathbf{H} + \mathbf{M} = \mu_0 \mathbf{H} + \mu_0 \chi_m \mathbf{H} = \mu_0 (1 + \chi_m) \mathbf{H} \quad (5.7)$$

Here we introduce the concept of relative permeability (μ_r).

$$\mu_r = 1 + \chi_m \quad (5.8)$$

By means of this the final equation can be written:

$$\mathbf{B} = \mu_0 \mu_r \mathbf{H} \quad (5.9)$$

5.3 Ampere's law and the magnetic material

In this section the vector calculus will be used at somewhat higher level.

The rotation operation (*rot*) generates a vector field which represents the vortexes of some vector field.

$$\mathbf{V}(\mathbf{r}) = V_x(x, y, z)\mathbf{i} + V_y(x, y, z)\mathbf{j} + V_z(x, y, z)\mathbf{k} \quad (5.10)$$

$$rot\mathbf{V}(\mathbf{r}) = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ V_x & V_y & V_z \end{vmatrix} \quad (5.11)$$

The Stoke's theorem integrates the rotation to a surface as follows:

$$\oint_g \mathbf{V}(\mathbf{r}) d\mathbf{r} = \oint_S (\text{rot} \mathbf{V}) d\mathbf{A} \quad (5.12)$$

Let us divide the following equation with μ_0 :

$$\mathbf{B} = \mu_0 \mathbf{H} + \mathbf{M} \quad (5.13)$$

$$\frac{\mathbf{B}}{\mu_0} = \mathbf{H} + \frac{\mathbf{M}}{\mu_0} \quad (5.14)$$

Generate the rotation of the equation above:

$$\text{rot}\left(\frac{\mathbf{B}}{\mu_0}\right) = \text{rot} \mathbf{H} + \text{rot}\left(\frac{\mathbf{M}}{\mu_0}\right) \quad (5.15)$$

Each term in the above equation can be interpreted separately based on the first Maxwell equation. The current densities are related both to the conductive current in the coil and to the magnetizing current.

$$\text{rot}\left(\frac{\mathbf{B}}{\mu_0}\right) = \mathbf{j}_{\text{tot}} \quad \text{rot} \mathbf{H} = \mathbf{j}_{\text{coil}} \quad \text{rot}\left(\frac{\mathbf{M}}{\mu_0}\right) = \mathbf{j}_{\text{magn}} \quad (5.16)$$

$$\mathbf{j}_{\text{tot}} = \mathbf{j}_{\text{coil}} + \mathbf{j}_{\text{magn}} \quad (5.17)$$

Stokes theorem generates integral form from the relations above:

$$\oint_g \left(\frac{\mathbf{B}}{\mu_0}\right) d\mathbf{r} = I_{\text{tot}} \quad \oint_g \mathbf{H} d\mathbf{r} = I_{\text{coil}} \quad \oint_g \left(\frac{\mathbf{M}}{\mu_0}\right) d\mathbf{r} = I_m \quad (5.18)$$

The central integral above is the well-known form of Ampere's law. This expresses that the curve integral of the \mathbf{H} vector on a closed path (g) equals the amount of the conductive current (current in a wire) surrounded by the g path. The integral on the right expresses that the curve integral the magnetization vector (\mathbf{M}/μ_0) equals the total magnetizing current surrounded by the g path. Finally the left hand side states that the curve integral of the magnetic induction vector (\mathbf{B}/μ_0) is equal with the total current (conductive and magnetizing) surrounded by the g path.

5.4 Inhomogeneous magnetic material

Consider two different magnetic materials with plane surface. The plane surfaces are connected thus creating an interface between the materials. This structure is subjected to the experimentation.

First the \mathbf{B} field is studied. The interface is contained by a symmetrical disc-like drum with the base area A . The upper and lower surface vectors are \mathbf{A}_1 and \mathbf{A}_2 respectively.

$$\mathbf{A}_1 = -\mathbf{A}_2 \quad |\mathbf{A}_1| = |\mathbf{A}_2| = A \quad (5.19)$$

Since magnetic monopoles do not exist the \mathbf{B} field does not have sources. So the flux of the \mathbf{B} field to a close surface is necessarily zero (Maxwell equation 3.)

$$\oint_S \mathbf{B} d\mathbf{A} = \mathbf{B}_1 \mathbf{A}_1 + \mathbf{B}_2 \mathbf{A}_2 = 0 \quad (5.20)$$

$$\mathbf{B}_1 \mathbf{A}_2 = \mathbf{B}_2 \mathbf{A}_2 \quad (5.21)$$

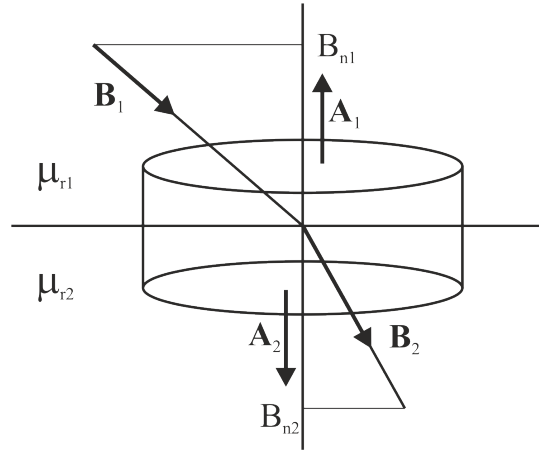


Figure 5.6: The \mathbf{B} field on the interface of different magnetic materials

The operation of dot product contains the projection of the \mathbf{B} vectors to the direction of \mathbf{A}_2 vector which is the normal direction to the surface. The subscript n means the absolute value of the normal direction component.

$$B_{1n} A_2 = B_{2n} A_2 \quad (5.22)$$

Once we are among real numbers the surface area cancels out readily.

$$B_{1n} = B_{2n} \quad (5.23)$$

According to this result the normal component of \mathbf{B} vector is continuous on the interface of magnetic materials.

Secondly the magnetic field (\mathbf{H}) is the subject of the analysis.

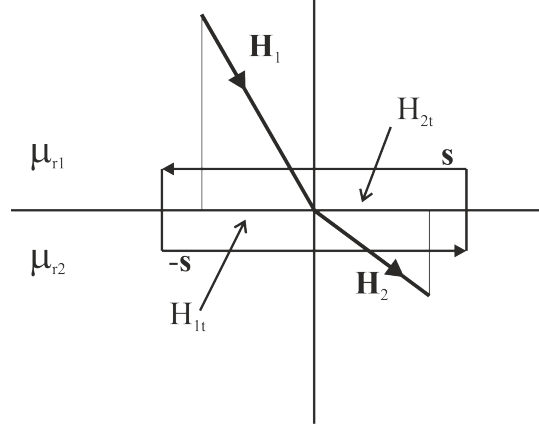


Figure 5.7: The \mathbf{H} field on the interface of different magnetic materials

The interface is surrounded by a very narrow rectangle-like loop with sections parallel and normal to the surface. The parallel sections of the loop are \mathbf{s} and $-\mathbf{s}$ vectors. The normal direction sections are ignored due to the infinitesimal size. The closed loop integral of the \mathbf{H} field equals the total conductive current contained by the loop according to Ampere's law. Here there is no such current so the right hand side of the equation will be zero.

$$\oint_g \mathbf{H} d\mathbf{r} = \mathbf{s} \mathbf{H}_1 + (-\mathbf{s}) \mathbf{H}_2 = 0 \quad (5.24)$$

$$\mathbf{s} \mathbf{H}_1 = \mathbf{s} \mathbf{H}_2 \quad (5.25)$$

The operation of dot product contains the projection of the \mathbf{H} vectors to the direction of \mathbf{s} vector which is the tangential direction to the surface. The subscript t means the absolute value of the tangential direction component.

$$s H_{1t} = s H_{2t} \quad (5.26)$$

Once we are among real numbers the length of the tangential section cancels out readily.

$$H_{1t} = H_{2t} \quad (5.27)$$

According to this result the tangential component of the \mathbf{H} vector is continuous on the interface of magnetic materials.

5.5 Demonstration example

A conductive rod with a radius ($R_1 = 10\text{cm}$) made of copper carries a uniform current density ($j = 10^5\text{A/m}^2$). The rod is surrounded by magnetic coating ($\mu_r = 10^3$) up to the radius ($R_2 = 15\text{cm}$). Find and sketch the radial dependence of H , B and M vectors. Determine the numerical peak values in the break points and find the amount of the magnetizing current.

First parameter to be calculated is the magnetic field (H). The tangential component of the H field is continuous on the interface of the magnetic materials. In our case the circular magnetic field is tangential to the surface of the magnetic coating, therefore the H field is unaffected by the presence of the coating.

In the rod

$$2r\pi \cdot H(r) = r^2\pi \cdot j$$

$$H(r) = \frac{j}{2}r$$

$$H(r = R_1) = \frac{j}{2}R_1 = \frac{10^5}{2}0.1 = 5 \cdot 10^3 \frac{\text{A}}{\text{m}}$$

$$I_{\text{Cond}} = 2R_1\pi \cdot H(r = R_1) = 3140\text{A}$$

Out of the rod

$$2r\pi \cdot H(r) = R_1^2\pi \cdot j$$

$$H(r) = \frac{j}{2} \frac{R_1^2}{r}$$

$$H(r = R_1) = \frac{j}{2}R_1 = \frac{10^5}{2}0.1 = 5 \cdot 10^3 \frac{\text{A}}{\text{m}}$$

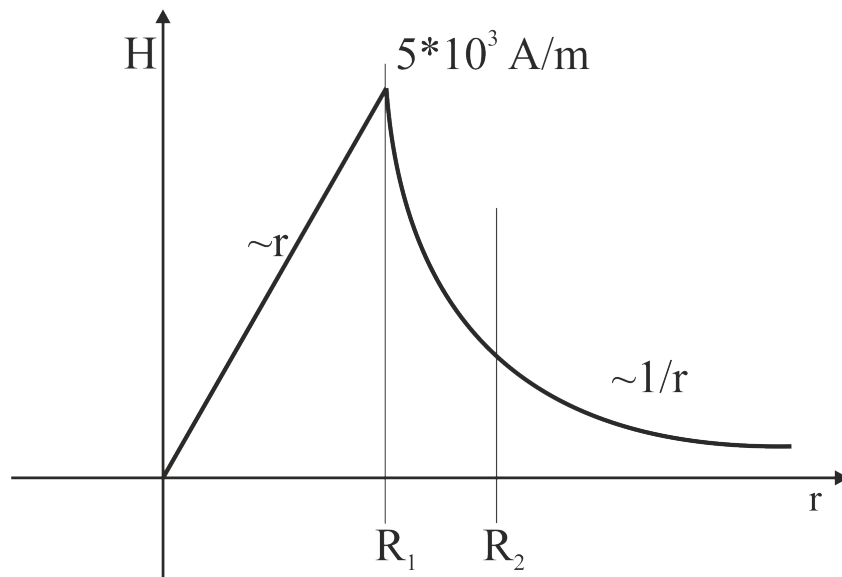


Figure 5.8: The magnetic field (H) as the function of distance

The peak value of the magnetic field (H) can be calculated as above. It shows that the function is continuous.

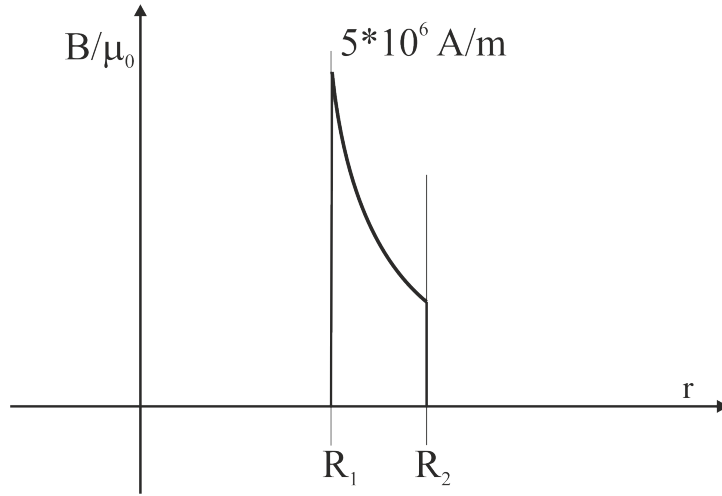


Figure 5.9: The magnetic induction (B/μ_0) as the function of distance

The B/μ_0 function is enlarged to μ_r times greater value where magnetic material is present. Difficulty lies in the drawing of the figure due to the huge ($\mu_r = 10^3$) multiplier. The peak value of B/μ_0 is $5 \cdot 10^6 \text{ A/m}$ at $r = R_1$ radius. The corresponding B value is 6.28 Tesla.

Magnetization results as the difference of the above functions:

$$\frac{\mathbf{M}}{\mu_0} = \frac{\mathbf{B}}{\mu_0} - \mathbf{H} \quad (5.28)$$

The peak value of M/μ_0 is $4.995 \cdot 10^6 \text{ A/m}$ at $r = R_1$ radius. The corresponding M value is 6.277 Tesla.

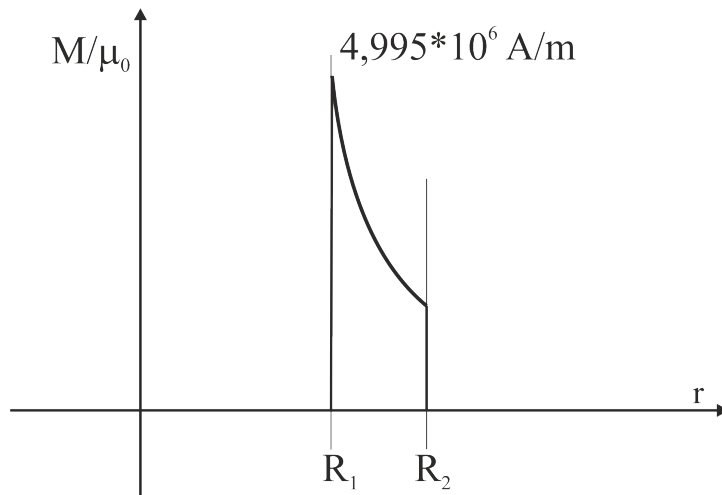


Figure 5.10: The magnetization (M/μ_0) as the function of distance

The amount of the magnetizing current can be calculated as follows:

$$\oint_g \left(\frac{\mathbf{M}}{\mu_0} \right) d\mathbf{r} = I_{magn} \quad (5.29)$$

The g curve of the integration is the circle with R_1 radius. The magnetizing current can be expressed as follows:

$$I_m = 2R_1\pi \frac{M}{\mu_0} = 2R_1\pi \left(\frac{B(R_1)}{\mu_0} - H(R_1) \right) = 2R_1\pi \left(\frac{1}{\mu_0} \mu_0 \mu_r \frac{j}{2} R_1 - \frac{j}{2} R_1 \right) = R_1^2 \pi \cdot j (\mu_r - 1) \quad (5.30)$$

The cross sectional area is multiplied with the current density. This is obviously the conductive current (I_{cond}) in the rod. The magnetizing current is as follows:

$$I_m = I_{Cond} \cdot (\mu_r - 1) = 3140 \cdot 999 = 3.14 \cdot 10^6 A \quad (5.31)$$

The magnetizing current is virtual current on the surface of the magnetic material. At R_1 radius the magnetizing current is in parallel direction with the conductive current while at R_2 radius the magnetizing current flows in opposite direction. At bigger radii than R_2 , the effects of two opposite direction magnetizing currents compensate each other so the magnetization intensity drops to zero.

5.6 Solenoid with iron core

In the former chapter at 5.2 section the empty solenoid coil has already been treated. Now the cavity contains the iron core which is characterized by μ_r value. The physical model of solenoid requires the length to be roughly ten times longer than the diameter.

Let us use Ampere's law. The magnetic field is homogeneous in the solenoid. A closed path is chosen which is parallel with the coil and located in the cavity on one side. The front side of the path is outside the coil where there is no magnetic field. The remaining two little sides of the rectangle are normal to the magnetic field thus can be ignored. The length is denoted l and the number of turns is N . So ultimately the Ampere's law emerges in a simple form:

$$\oint_{loop} \mathbf{H}(\mathbf{r}) d\mathbf{r} = \sum I \quad (5.32)$$

$$Hl = NI \quad (5.33)$$

$$H = \frac{NI}{l} \quad (5.34)$$

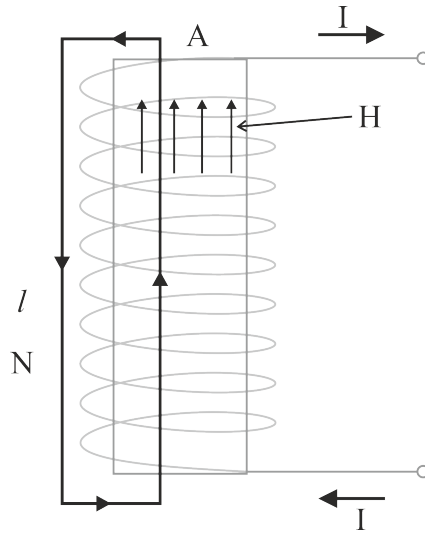


Figure 5.11: Solenoid with core

The generated magnetic induction is as follows:

$$B = \mu_0 \mu_r \frac{NI}{l} \quad (5.35)$$

The generated turn flux and coil flux values can be expressed:

$$\Phi_{turn} = BA = \frac{\mu_0 \mu_r NIA}{l} \quad \Phi_{coil} = BA \cdot N = \frac{\mu_0 \mu_r N^2 IA}{l} \quad (5.36)$$

5.7./ Toroid coil with air gap in the iron core

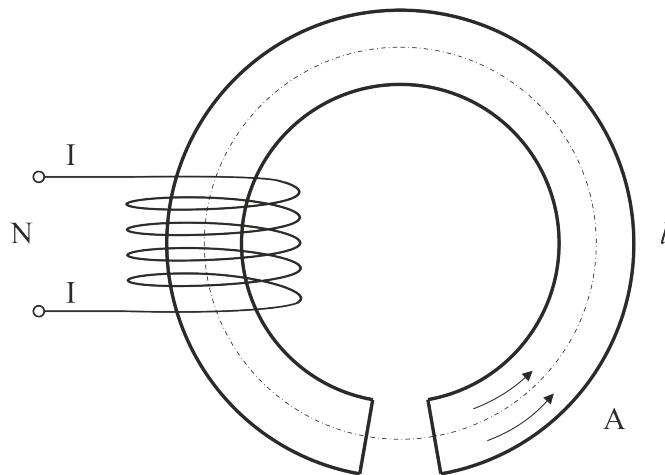


Figure 5.12: Toroid with air gap

The air gap is perpendicular to the magnetic field in the coil therefore the magnetic induction is continuous on the surfaces of the air gap since the normal component of the B field is always continuous on the surface of magnetic material.. That means the B magnetic induction is constant all around the coil.

$$B_{iron} = B_{air} = B \quad H_{air} = \frac{B}{\mu_0} \quad H_{iron} = \frac{B}{\mu_0 \mu_r} \quad (5.37)$$

Let us use the Ampere's law:

$$H_{iron}l + H_{air}\delta = NI \quad (5.38)$$

The letters l and δ are the circumference of the coil and width of the air gap respectively.

After substitution:

$$\frac{B}{\mu_0 \mu_r}l + \frac{B}{\mu_0}\delta = NI \quad (5.39)$$

From here B can readily be expressed:

$$B = \frac{\mu_0 NI}{\frac{l}{\mu_r} + \delta} \quad (5.40)$$

The generated turn flux and coil flux values can be expressed:

$$\Phi_{turn} = BA = \frac{\mu_0 NIA}{\frac{l}{\mu_r} + \delta} \quad \Phi_{coil} = BA \cdot N = \frac{\mu_0 N^2 IA}{\frac{l}{\mu_r} + \delta} \quad (5.41)$$

Chapter 6

Time dependent electromagnetic field - György Hárs

The phenomena of electromagnetic induction are majorly important in the applications. The production and the transformation of electric energy are carried out this way.

6.1 Motion related electromagnetic induction

6.1.1 Plane generator (DC voltage)

The plane generator is a hypothetical device which is unpractical to use in its original form, however it is capable of demonstrating the operation some practically used generators. The physical principles of operation are clearly apparent without the disturbing technical details.

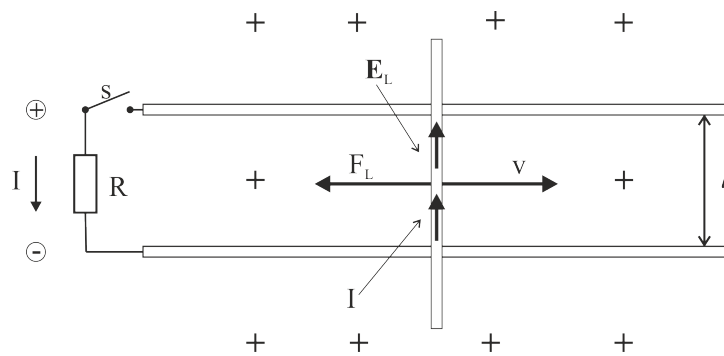


Figure 6.1: Plane generator

The plane generator consists of two parallel conductive rails and a similarly conductive crossbar perpendicular to the rails. The crossbar can travel freely on the rails while

staying in galvanic contact with both. A resistor of resistance R is also connected between the rails. The whole setup is placed into homogeneous B magnetic field which points into the plane of the paper. Let us move the crossbar with uniform v velocity parallel with the rails. Let the velocity vector point to the right hand side direction.

In the first part of the experiment the S switch which connects the resistor to the setup is open.

The generated Lorentz electric field is as follows:

$$\mathbf{E}_L = \mathbf{v} \times \mathbf{B} \quad (6.1)$$

According to the vector product the Lorentz electric field points upside direction. The Lorentz electric field pushes the positive charge carriers to upside direction so the upside terminal is the positive one. The result is the same if electrons are considered as charge carriers. This time the electrons are pushed downside direction making the downside terminal negative which matches the earlier result.

The \mathbf{v} and \mathbf{B} vectors are normal to each other so the absolute value of the result is merely the product of the absolute values:

$$E_L = v \cdot B \quad (6.2)$$

The absolute value of the induced voltage can be calculated without any integration by a simple product.

$$U_{ind} = E_L \cdot l = B \cdot l \cdot v \quad (6.3)$$

Here the distance between the rails is denoted l .

The absolute value of the induced voltage can also be calculated in the following way: The magnetic flux affecting the setup is the product of the magnetic induction (B) and the active area (A).

$$\Phi = BA = Blx \quad (6.4)$$

The time derivative of the above formula is as follows:

$$\frac{d\Phi}{dt} = B \frac{dA}{dt} = Bl \frac{dx}{dt} = B \cdot l \cdot v \quad (6.5)$$

The result matches the absolute value of the induced voltage. The polarity of the result should be considered separately. If the velocity vector points to the right hand side direction then the active flux increases due to the increasing area. The increasing flux points into the paper so the generated electric field supposed to show a clockwise rotation. In contrast to this the rotation of the electric field is counter clockwise as it has been

shown in the first part of this argument. Altogether one can summarize the conclusion in the following formula:

$$U_{ind} = -\frac{d\Phi}{dt} \quad (6.6)$$

This formula is the famous Faraday induction law. This is true for all kinds of induction processes.

Now let us return to the discussion of the plane generator by closing the S switch, this way applying a load to the generator. Current flowing through the resistor is denoted i .

$$i = \frac{U_{ind}}{R} = \frac{Blv}{R} \quad (6.7)$$

The electric power generated P_{el} is the following:

$$P_{el} = U_{ind}i = \frac{(Blv)^2}{R} \quad (6.8)$$

The induced current flows through the crossbar. The current and magnetic field interact according to Lorentz law.

$$\mathbf{F}_L = i\mathbf{l} \times \mathbf{B} \quad (6.9)$$

The direction of the Lorentz force is just the opposite of the velocity. If I move the crossbar with my hand I have to overcome the Lorentz force.

This is the point to mention Lenz's law which states that following: The direction of the induced current is determined accordingly, that by means of its magnetic field, the induced current always opposes the original change in the magnetic flux. So if the flux is increased by my hand, the induced current opposes my hand's motion.

My force will be parallel direction with the velocity. This way I make positive power on the system.

$$F_L = il \cdot B = \frac{Blv}{R} l \cdot B \quad (6.10)$$

The positive power exerted to the system is the product of the force and the velocity:

$$P_{mech} = F_L \cdot v = \frac{(Blv)^2}{R} \quad (6.11)$$

The amount of the electrical power matches the formula of the mechanical power. This means that the mechanical power required to move the crossbar against the force of the magnetic field is equal to the electrical power which heats up the resistor.

Based on the principle of the plane generator there are practically usable generator types such as the "Drum generator" and the "Unipolar generator" both generating DC voltage.

6.1.2 Rotating frame generator (AC voltage)

The rotating frame generator is a hypothetical device which is unpractical to use in its original form, however it is capable of demonstrating the operation some practically used generators. The physical principles of operation are clearly apparent without the disturbing technical details.

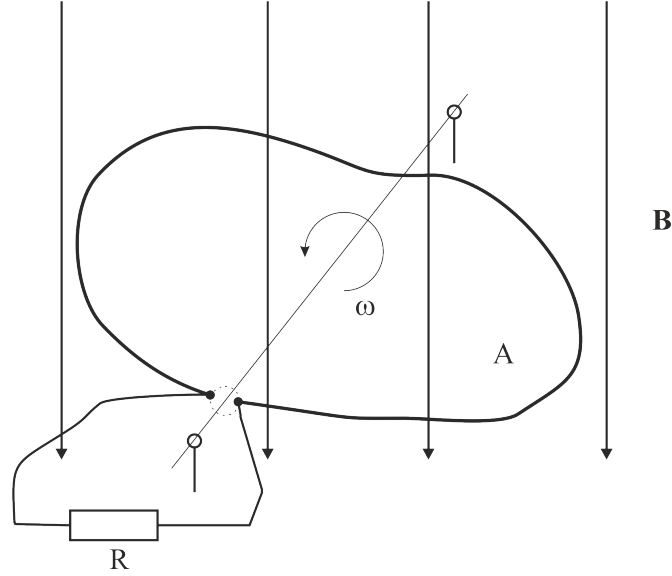


Figure 6.2: Rotating frame generator

The rotating frame generator consists of some kind of wire frame with the surface area A without any special condition for the shape of the frame. The frame is rotating around an axis which is expanded between two diagonal points of the frame. The axis is positioned perpendicular in a homogeneous magnetic field. On the rotational axis there is a pair of sliding rings which is solidly connected to the two ends of the cut wire frame. The sliding connectors are hooked up to a resistor of resistance R through an S switch which is open during the first part of the experiment.

The flux in the frame as the function of time can be written easily:

$$\Phi(t) = BA \cos(\omega t) \quad (6.12)$$

Let us use the Faraday induction law:

$$U_{ind} = -\frac{d\Phi}{dt} = BA\omega \sin(\omega t) \quad (6.13)$$

The induced current is expressed by Ohm's law:

$$i = \frac{U_{ind}}{R} = \frac{BA\omega}{R} \sin(\omega t) \quad (6.14)$$

The electrical power generated is as follows:

$$P_{el} = U_{ind}i = \frac{(BA\omega)^2}{R} \sin^2(\omega t) \quad (6.15)$$

Let us check out the mechanical power required to rotate the generator.

The torque \mathbf{M} is affecting a magnetic dipole in a \mathbf{B} magnetic field. It has already been discussed in chapter 4.

$$\mathbf{M} = i\mathbf{A} \times \mathbf{B} \quad (6.16)$$

The absolute value of the torque is as follows:

$$M = iAB \sin(\omega t) \quad (6.17)$$

The mechanical power exerted to the system is the product of the torque and the angular velocity of the rotation.

$$P_{mech} = M\omega = \frac{BA\omega}{R} \sin(\omega t) \cdot AB \sin(\omega t) \cdot \omega = \frac{(BA\omega)^2}{R} \sin^2(\omega t) \quad (6.18)$$

The final formula of the mechanical power completely matches the formula of the electrical consumption. So altogether the situation is clear. The torque of my hand which rotates the frame generator overcomes the opposition of the induced current according to Lenz's law. The generated power has been consumed in the resistor by warming it up.

Generators which supply AC voltage into the electrical energy systems operate on the principle of the rotating frame generator.

6.1.3 Eddy currents

If the magnetic field changes over time inside of a conductive medium, the generated electric field gives rise to loop currents which circulate in the medium. These are the eddy currents which cause energy dissipation in the medium. The direction of the current is determined by Lenz's law.

Swinging rings experiment

The rings are made of aluminum with an approximate diameter of twenty centimeters. One of them has a thin cutting so this ring is not continuous all around. The rings are suspended according to the figure. A bar magnet is pushed back and forth into the ring. The ring with the cutting is unaffected by the periodic motion of the magnet. However the intact ring gradually starts to swing if the operator moves the magnet in synchronism with the oscillation frequency of the pendulum.

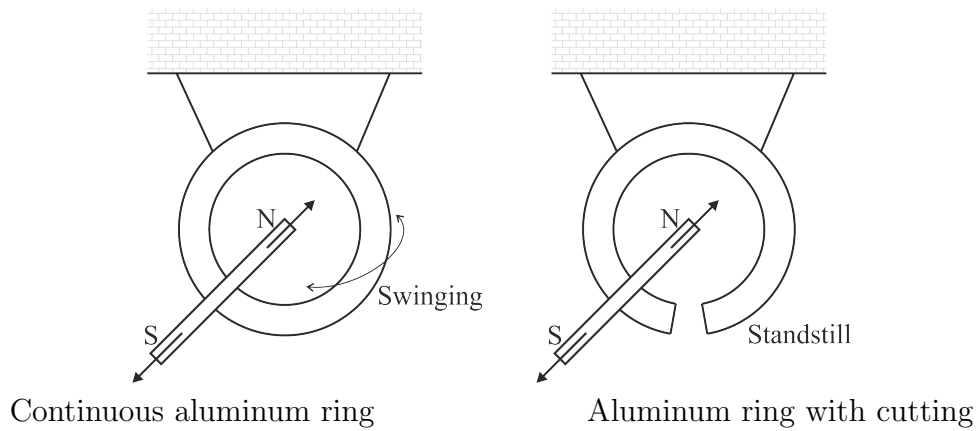


Figure 6.3: Swinging rings experiment

Explanation: The motion of the bar magnet causes the variation of the flux in the ring. The induced electric field generates the induced loop current (eddy current). The magnetic field of the induced loop current opposes the original effect according to Lenz's law. The original effect is the motion of the bar magnet which can not be stopped, therefore the ring starts to swing by the periodic effect of the braking force. If the ring with the cutting is taken, no effect will show up, since the cutting inhibits the formation of the eddy current.

Waltenhofen pendulum

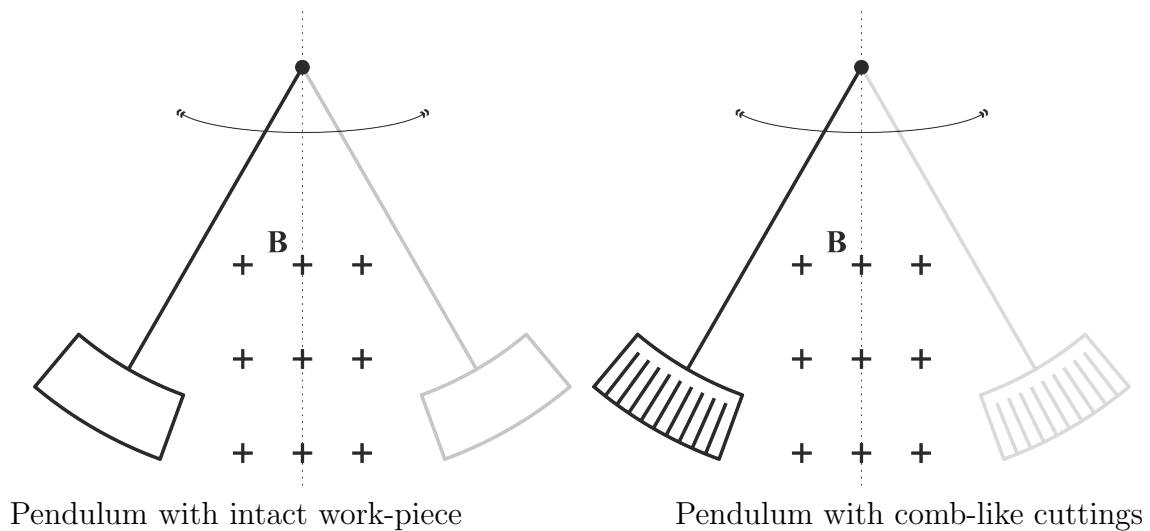


Figure 6.4: Waltenhofen pendulum experiment

A pendulum is made with an aluminum work piece on its end according to the figure. One of the work-pieces is intact the other one is having comb-like cuttings. The intact work-piece is swinging between the jaws of the electromagnet which is inactive. Let the pendulum swing and observe that the attenuation of the motion is insignificant. Now switch DC voltage to the electromagnet. The swinging will stop completely in three oscillations. Now replace the work-piece for that with the cuttings. By repeating the experiment the attenuation does not appear on switching the magnet.

Explanation: When the intact piece was moving through the magnetic field, eddy current was generated in the work piece by the effect of the flux variation. The magnetic effect of the eddy current attenuated the swing according to Lenz's law. Once the work-piece with cuttings has been installed the attenuation failed since eddy currents have been prevented from happening.

6.2 Electromagnetic induction at rest

Electromagnetic induction can also occur without mechanical motion. The primary cause of the induction process is the variation of the electric current, which in turn generates time variant magnetic field.

6.2.1 The mutual and the self induction

Consider n pieces of current loops. Each of them carries a current i_i and each of them contains a flux Φ_i . The flux is originated partly from its own current and partly by all other current loops.

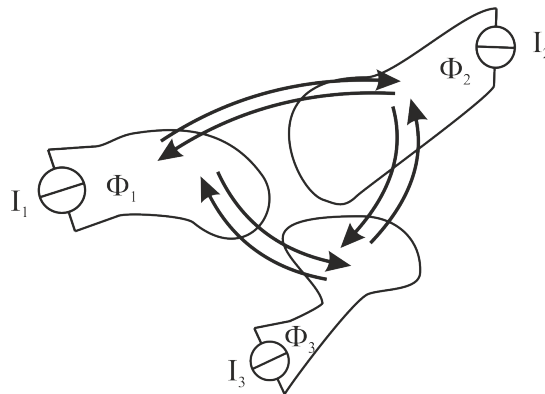


Figure 6.5: Current loops

Experience shows that the coupled fluxes and the self fluxes are proportional with the

corresponding currents so the total flux in a loop can be expressed by a linear relation.

$$L_1 i_1 + M_{12} i_2 + M_{13} i_3 = \Phi_1 \quad (6.19)$$

$$M_{21} i_1 + L_2 i_2 + M_{23} i_3 = \Phi_2 \quad (6.20)$$

$$M_{31} i_1 + M_{32} i_2 + L_3 i_3 = \Phi_3 \quad (6.21)$$

This relation is presented in the best way by using matrix formalism.

$$\begin{bmatrix} L_1 & M_{12} & M_{13} \\ M_{21} & L_2 & M_{23} \\ M_{31} & M_{32} & L_3 \end{bmatrix} \cdot \begin{bmatrix} i_1 \\ i_2 \\ i_3 \end{bmatrix} = \begin{bmatrix} \Phi_1 \\ \Phi_2 \\ \Phi_3 \end{bmatrix} \quad (6.22)$$

Here the self induction coefficients and the mutual induction coefficients are denoted with letters L and M respectively. In the M coefficients the first subscripts indicate the current loop that received the external flux while the second subscript shows the current loop that generated the magnetic field. Obviously the self induction coefficients do not need double subscripts.

The major physical point of the above description is the fact that the induction matrix is symmetrical. That means for instance M_{12} and M_{21} elements are equal. This contains the important fact that by applying current to loop 1 and measuring the flux in loop 2 the mutual induction coefficient turns out to be the same when the current loop and the measuring loop are swapped. This symmetry provides the possibility to determine the mutual induction coefficient in the most convenient way.

6.2.2 Induced voltage of a current loop

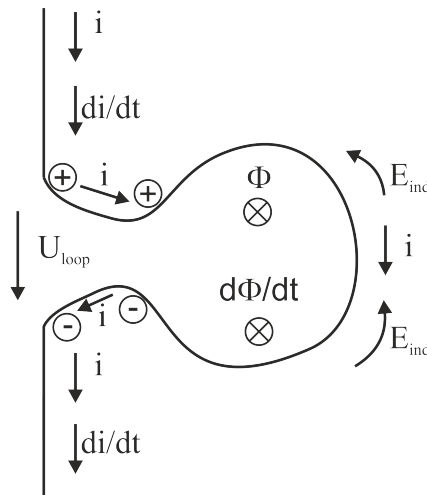


Figure 6.6: Induced voltage on current loop

Consider a current loop according to the figure above. Let us assume an electric current flowing down-side direction and increasing in magnitude. Therefore the time derivative of the current also points down-side direction. In the loop, both the generated flux and the time derivative of the flux point into the sheet of the paper. The induced electric field (some earlier books call it as electromotive force) in the loop performs a counter clockwise rotation due to the negative sign in the Faraday induction law.

$$U_{ind} = -\frac{d\Phi}{dt} \quad (6.23)$$

Accordingly, the induced electric field pushes the positive charge carriers counter clockwise direction. This means that the positive pole of the induced voltage will be the upper pole and the negative is the lower one. The voltage drop of a two-pole component is directed from the positive to the negative pole. So the measurable loop voltage is pointing down-side direction, similarly to the direction of the time derivative of the current. Thus the measured loop voltage on the two-pole component will be the time derivative of the loop current multiplied with the self induction coefficient without the negative sign.

$$U_{loop} = L \frac{di}{dt} \quad (6.24)$$

If there are more loops in the setup the loop voltages can be expressed as the time derivative of the above matrix equation:

$$\begin{bmatrix} L_1 & M_{12} & M_{13} \\ M_{21} & L_2 & M_{23} \\ M_{31} & M_{32} & L_3 \end{bmatrix} \cdot \begin{bmatrix} di_1/dt \\ di_2/dt \\ di_3/dt \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} \quad (6.25)$$

6.2.3 The transformer

The transformer consists of at least two coils on a common iron core. The coil which receives the excitation is called the primary coil, while the coil which provides the output is called the secondary coil. Two geometrical arrangements will be discussed, the solenoid and the toroid.

These two types of geometry have already been discussed in chapter 5 in some extent. The discussion of these types will be carried out in a uniform way with the following notations: The cross sectional area of the coil is A . The length of the solenoid and the circumference of the toroid are denoted l . The number of turns is denoted N_1 and N_2 . In both cases the Ampere's law emerges in a simple form:

$$\oint_{loop} \mathbf{H}(\mathbf{r}) d\mathbf{r} = \sum I \quad (6.26)$$

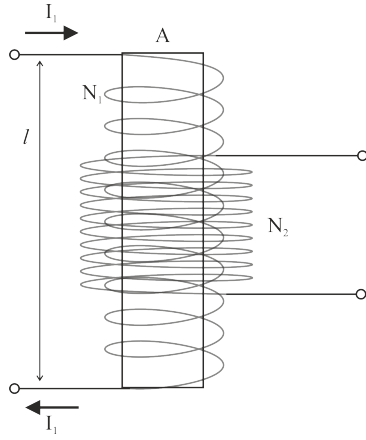


Fig. 6.7 The solenoid transformer

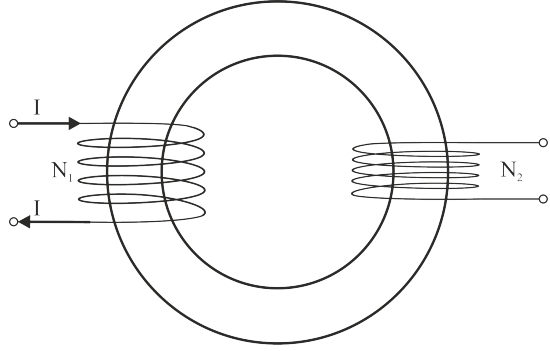


Fig. 6.8 The toroid transformer

$$Hl = N_1 I \quad (6.27)$$

$$H = \frac{N_1 I}{l} \quad (6.28)$$

The generated magnetic induction is as follows:

$$B = \mu_0 \mu_r \frac{N_1 I}{l} \quad (6.29)$$

The generated turn flux can be expressed:

$$\Phi_{turn} = BA = \frac{\mu_0 \mu_r N_1 I A}{l} \quad (6.30)$$

The voltage of one turn of the coil is called the turn voltage. This is the time derivative of the turn flux.

$$U_{turn} = \frac{dB}{dt} A = \frac{\mu_0 \mu_r N_1 A}{l} \frac{dI}{dt} \quad (6.31)$$

The coil voltage on the primary coil is as follows:

$$U_1 = N_1 \cdot U_{turn} = \frac{\mu_0 \mu_r N_1^2 A}{l} \frac{dI}{dt} = L_1 \frac{dI}{dt} \quad (6.32)$$

The formula of the self induction coefficient can be identified:

$$L_1 = \frac{\mu_0 \mu_r N_1^2 A}{l} \quad (6.33)$$

Similarly for the secondary coil the self induction coefficient is as follows:

$$L_2 = \frac{\mu_0 \mu_r N_2^2 A}{l} \quad (6.34)$$

The coil flux can also be expressed:

$$\Phi_{1coil} = N_1 \Phi_{turn} = L_1 \cdot I \quad (6.35)$$

The secondary voltage also can be determined by means of the turn voltage:

$$U_2 = N_2 \cdot U_{turn} = \frac{\mu_0 \mu_r N_1 N_2 A}{l} \frac{dI}{dt} = M \frac{dI}{dt} \quad (6.36)$$

The formula of the mutual induction coefficient can be identified:

$$M = \frac{\mu_0 \mu_r N_1 N_2 A}{l} \quad (6.37)$$

An important relation can be seen easily:

$$M^2 = L_1 L_2 \quad (6.38)$$

Let us compare the primary and the secondary voltages.

$$\frac{U_2}{U_1} = \frac{N_2 U_{turn}}{N_1 U_{turn}} = \frac{N_2}{N_1} = T \quad (6.39)$$

The ratio of number of turns is called the transmission (T) of the transformer.

So far there was no any load on the secondary coil. Let us analyze the case when load is applied to the secondary coil. The analysis is carried out by means of the complex amplitudes of the sinusoidal quantities. Here j denotes the imaginary unit

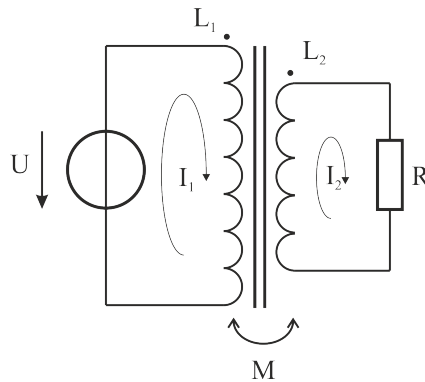


Figure 6.9: Transformer with load

The Kirchhoff loop equations are used.

$$j\omega L_1 I_1 - j\omega M I_2 - U = 0 \quad (6.40)$$

$$R I_2 + j\omega L_2 I_2 - j\omega M I_1 = 0 \quad (6.41)$$

Regroup the equation.

$$j\omega L_1 I_1 - j\omega M I_2 = U \quad (6.42)$$

$$-j\omega M I_1 + j\omega L_2 I_2 = -R I_2 \quad (6.43)$$

The first and the second equations are multiplied with M and L_1 respectively.

$$j\omega L_1 M I_1 - j\omega M^2 I_2 = U M \quad (6.44)$$

$$-j\omega L_1 M I_1 + j\omega L_1 L_2 I_2 = -R L_1 I_2 \quad (6.45)$$

Add the equations together. The first term cancels out.

$$j\omega I_2 (L_1 L_2 - M^2) = U M - R L_1 I_2 \quad (6.46)$$

$$I_2 (j\omega (L_1 L_2 - M^2) + R L_1) = U M \quad (6.47)$$

From here I_2 can be expressed:

$$I_2 = \frac{U}{j\omega \left(\frac{L_1 L_2}{M} - M \right) + \frac{R L_1}{M}} = \frac{U}{j\omega (M - M) + \frac{R_1}{T}} = \frac{U T}{R} \quad (6.48)$$

Here $L_1 L_2 = M^2$ and *the* $L_1/M = 1/T$ relations have been used.

The generated secondary voltage is simply:

$$U_2 = R I_2 = U T \quad (6.49)$$

This equation proves that the secondary voltage matches the value of the case without load. This result is the conclusion of the model we are using in which the coils are free of serial resistance. The output effective power is the half value of the product between the voltage and the current.

$$P_{out} = \frac{1}{2} U_2 I_2 = \frac{(U T)^2}{2 R} \quad (6.50)$$

Now let us find out the relations on the primary side. For this purpose the value of I_1 should be determined from the initial equations.

$$j\omega L_1 I_1 - j\omega M I_2 = U \quad (6.51)$$

The value of I_2 is substituted:

$$j\omega L_1 I_1 - j\omega M \frac{UT}{R} = U \quad (6.52)$$

I_1 can be expressed readily.

$$I_1 = U \frac{1 + j\omega \frac{M}{R} T}{j\omega L_1} = \frac{U}{R} \frac{R + j\omega MT}{j\omega L_1} = \frac{U}{R} \left(\frac{R}{j\omega L_1} + \frac{MT}{L_1} \right) = \frac{U}{R} \left(\frac{R}{j\omega L_1} + T^2 \right) \quad (6.53)$$

$$I_1 = \frac{U}{j\omega L_1} + \frac{U}{R} T^2 \quad (6.54)$$

Two conclusions can be drawn.

The first one is related to the input power. The input current above consists of two terms. The first one is a reactant current which is in ninety degree lag relative to the voltage. This term does not produce effective power. The second term is in phase with the voltage so the effective power will be generated accordingly.

$$P_{in} = \frac{1}{2} U I_1 = \frac{(UT)^2}{2R} \quad (6.55)$$

The result perfectly matches the form of the output power.

The second conclusion refers to the input impedance of the transformer. Let us calculate the Z_{in} value which is the ratio of the input voltage over the primary current.

$$Z_{in} = \frac{U}{I_1} = \frac{U}{\frac{U}{j\omega L_1} + \frac{U}{R} T^2} = \frac{1}{\frac{1}{j\omega L_1} + \frac{1}{R} T^2} = \frac{R}{\frac{R}{j\omega L_1} + T^2} \quad (6.56)$$

Provided the frequency is high enough, the first term in the denominator can be ignored relative to the square of the transmission. This time the input impedance is real resistance.

$$R_{in} = R \left(\frac{1}{T} \right)^2 \quad (6.57)$$

The second conclusion is the fact that the load resistance on a transformer shows up on the input of the transformer as a real resistance with the value above. So the rule of thumb can be declared, that the load is transformed to the input with the square of the transmission.

6.2.4 Energy stored in the coil

Let us increase the current in a solenoid coil from zero to some I value gradually in time. The coil reacts with an opposition to the increasing current according to Lenz's law. The induced voltage of the coil needs to be overcome in order to press through the current. This way we have to carry out positive work and finally the coil will possess magnetic energy. The induced voltage is expressed by the known formula:

$$U = L \frac{dI}{dt} \quad (6.58)$$

Multiplying it with the current one recovers to invested power.

$$P(t) = UI = (L \frac{dI}{dt})I = L(I \frac{dI}{dt}) \quad (6.59)$$

On the right hand side a function and its time derivative are multiplied together. It is known from math that the following rule is true.

$$f(x) \frac{df(x)}{dx} = \frac{1}{2} \frac{d}{dx} f^2(x) \quad (6.60)$$

Let us apply this rule to the original case.

$$I \frac{dI}{dt} = \frac{1}{2} \frac{d}{dt} (I^2) \quad (6.61)$$

This can be readily substituted.

$$P(t) = \frac{1}{2} L \frac{d}{dt} (I^2(t)) \quad (6.62)$$

Let us integrate the two sides of the equation over time with homogeneous initial conditions. Accordingly in the initial state there was no current and no energy stored in the coil.

$$\int_0^t P(t') dt' = \frac{L}{2} \int_0^t \frac{d}{dt'} (I^2(t')) dt' = \frac{L}{2} \int_0^t d(I^2(t')) = \frac{L}{2} I^2(t) \quad (6.63)$$

The integral of the invested power is the magnetic energy stored. $\int_0^t P(t') dt' = E_m(t)$

The last two equations combined provide the final result of the magnetic energy of the coil:

$$E_m = \frac{1}{2} L I^2 \quad (6.64)$$

Let us substitute the formula of the self induction coefficient.

$$L = \frac{\mu_0 \mu_r N^2 A}{l} \quad E_m = \frac{L}{2} I^2 = \frac{1}{2} \frac{\mu_0 \mu_r N^2 A}{l} I^2 \quad (6.65)$$

The magnetic energy density (ε_m) in the coil is the ratio of the energy over the volume (Al) :

$$\varepsilon_m = \frac{E_m}{Al} = \frac{1}{2} \frac{\mu_0 \mu_r N^2}{l^2} I^2 = \frac{1}{2} \mu_0 \mu_r \left(\frac{NI}{l} \right)^2 \quad (6.66)$$

The following two formulas are well-known:

$$\frac{NI}{l} = H \quad \mu_0 \mu_r H = B \quad (6.67)$$

By means of these, the magnetic energy density in the coil comes out.

$$\varepsilon_m = \frac{1}{2} HB \quad \varepsilon_m = \frac{1}{2} \mathbf{H} \mathbf{B} \quad \left[\frac{W}{m^3} \right] \quad (6.68)$$

Though this result was deduced for the specific condition of a solenoid coil, the formula for energy density is universally valid for any condition and geometry. In general case the dot product of the magnetic field and the magnetic induction provides the result.

6.3 The Maxwell equations

In the table below the Maxwell equations are summarized along with some auxiliary relations between the field parameters.

Maxwell 1. Ampere's law This expresses the fact that the magnetic force lines do not have starting and final points, much rather they are closing into themselves like closed loops. The rotation of the magnetic force lines is determined by the sum of the conductive current and the displacement current. The displacement current also generates magnetic field without any electric conduction. This makes possible the electromagnetic wave propagation in the space.

Maxwell 2. Faraday induction law This law states that the changing magnetic field generates circulating electric field around itself. The direction of the circulation is opposite of the right hand screw direction. In absence of magnetic field the electric field is circulation free, thus scalar potential can be introduced.

Maxwell 3. Magnetic Gauss law This equation declares that magnetic monopoles do not exist, so the magnetic flux to any closed surface is zero.

Maxwell 4. Gauss law This law declares that the electric force lines start on the positive charge and end on the negative charge. The electric flux of on a closed surface is proportional to the amount of the contained free charges and it is zero if the charges are out of the closed volume.

Stokes' and Gauss Ostrogradsky theorems provide the conversion between the differential and integral laws.

Equation number	Differential form	Integral form
Maxwell 1. Ampere's law	$rot\mathbf{H} = \mathbf{j} + \frac{\partial \mathbf{D}}{\partial t}$	$\oint_g \mathbf{H} d\mathbf{r} = I_{cond} + \frac{d\Phi_D}{dt} \quad \Phi_D = \int_S \mathbf{D} d\mathbf{A}$ $\varepsilon_0 \mu_0 = c^2 \quad \mathbf{D} = \varepsilon_0 \varepsilon_r \mathbf{E}$
Maxwell 2. Faraday induction law	$rot\mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$	$\oint_g \mathbf{E} d\mathbf{r} = -\frac{d\Phi_B}{dt} \quad \Phi_B = \int_S \mathbf{B} d\mathbf{A}$ $\varepsilon_0 \mu_0 = c^2 \quad \mathbf{B} = \mu_0 \mu_r \mathbf{H}$
Stokes' theorem	Conversion $\oint_g \mathbf{v} d\mathbf{r} = \int_S rot\mathbf{v} \cdot d\mathbf{A}$	
Maxwell 3 Magnetic Gauss law	$div\mathbf{B} = 0$	$\oint_V \mathbf{B} d\mathbf{A} = 0$
Maxwell 4. Gauss law	$div\mathbf{D} = \rho_{free}$	$\oint_V \mathbf{D} d\mathbf{A} = Q_{free}$
Gauss Ostrogradsky theorem	Conversion $\oint_S \mathbf{v} d\mathbf{A} = \oint_V div\mathbf{v} \cdot dV$	

Chapter 7

Electromagnetic oscillations and waves - Gábor Dobos

7.1 Electrical oscillators

In the previous semester we have discussed mechanical oscillators already. Although the physical processes in an electrical oscillator are very different from those in a mechanical oscillator, the equations describing them take similar forms, and result in similar behaviour. The simplest electrical oscillator consists of an ideal inductor and capacitor. The voltage of the inductor is:

$$U_L = L \frac{dI}{dt} \quad (7.1)$$

where I is the current running through the inductor, and L is its self-inductance, respectively. The voltage of the capacitor is proportional to the charge accumulated in it:

$$U_C = \frac{Q}{C} = \frac{1}{C} \int I dt \quad (7.2)$$

where Q is the accumulated charge in the capacitor, C is its capacitance and I is the current running into it. A simple circuit consisting of only these two elements can be described using Kirchoff's second rule:

$$\sum_i U_i = 0 \quad (7.3)$$

$$L \frac{dI}{dt} + \frac{Q}{C} = 0 \quad (7.4)$$

$$L \frac{d^2 I}{dt^2} + \frac{1}{C} \frac{dQ}{dt} = 0 \quad (7.5)$$

$$L \frac{d^2 I}{dt^2} + \frac{1}{C} I = 0 \quad (7.6)$$

(7.6) has the same form, as the equation, describing the undamped mechanical oscillator ($m \frac{d^2 x}{dt^2} + Dx = 0$), with x replaced by I , m replaced by L and D replaced by $\frac{1}{C}$, respectively. As the two equations are identical, so are their solutions:

$$I(t) = I_0 \cos(\omega_0 t + \phi_0) \quad (7.7)$$

where $\omega_0 = \sqrt{\frac{1}{LC}}$, and I_0 and ϕ_0 are determined by the initial conditions.

In case of a mechanical oscillator the viscosity of the medium in which the oscillator is moving caused damping. In case of an electrical oscillator the ohmic resistance of the electrical components has a similar effect. A circuit consisting of an inductor, a resistor and a capacitor connected in a loop behaves as a damped oscillator:

$$\sum_i U_i = 0 \quad (7.8)$$

$$L \frac{dI}{dt} + RI + \frac{Q}{C} = 0 \quad (7.9)$$

$$L \frac{d^2 I}{dt^2} + R \frac{dI}{dt} + \frac{1}{C} \frac{dQ}{dt} = 0 \quad (7.10)$$

$$L \frac{d^2 I}{dt^2} + R \frac{dI}{dt} + \frac{1}{C} I = 0 \quad (7.11)$$

Introducing the constant $\beta = \frac{R}{2L}$:

$$\frac{d^2 I}{dt^2} + 2\beta \frac{dI}{dt} + \omega_0^2 I = 0 \quad (7.12)$$

The equation has three different solutions depending on ω_0 and β :

- In the underdamped case ($\omega_0 > \beta$)

$$I_u(t) = I_0 e^{-\beta t} \cos(\omega t + \phi_0) \quad (7.13)$$

where $\omega = \sqrt{\omega_0^2 - \beta^2}$

- In the critically damped case ($\omega_0 = \beta$)

$$I_c(t) = (A + Bt) e^{-\beta t} \quad (7.14)$$

- In the overdamped case ($\omega_0 < \beta$)

$$I_o(t) = I_1 e^{-\lambda_1 t} + I_2 e^{-\lambda_2 t} \quad (7.15)$$

Damped electrical oscillators lose energy due to losses on the ohmic resistance of electrical components. Even in the underdamped case the amplitude of the oscillation decreases exponentially, and the frequency is lower than that of the undamped oscillator. If the ohmic resistance (the damping) is increased, the period of the oscillation also increases. When $\frac{R}{2L} = \sqrt{\frac{1}{LC}}$, the period reaches infinity: even a single cycle of the oscillation would take an infinitely long time. For higher ohmic resistances, no oscillations are possible.

Since all practical electronic components have some ohmic resistance, all practical electrical oscillators behave as damped oscillators. Even if the damping is weak, the oscillation continuously decays due to ohmic losses. Similarly to mechanical oscillators they require external power to function for extended periods of time. This can be achieved by connecting a voltage source into the circuit:

$$L \frac{dI}{dt} + RI + \frac{Q}{C} = U_0 \sin(\omega_f t) \quad (7.16)$$

$$L \frac{d^2 Q}{dt^2} + R \frac{dQ}{dt} + \frac{1}{C} Q = U_0 \sin(\omega_f t) \quad (7.17)$$

$$\frac{d^2 I}{dt^2} + \frac{R}{L} \frac{dI}{dt} + \frac{1}{LC} Q = \frac{U_0}{L} \sin(\omega_f t) \quad (7.18)$$

This equation is identical to the equation describing the forced mechanical oscillator, and its solution also takes the same form. The solution is the sum of the general solution of the homogenous equation plus a particular solution of the inhomogeneous equation. Similarly to mechanical oscillators, the homogenous solution decays exponentially, and after a short transient only the inhomogenous solution remains:

$$Q(t) = Q_0 \cos(\omega_f t + \Delta\phi) \quad (7.19)$$

Where

$$Q_0 = \frac{U_0/L}{\sqrt{(\omega_0^2 - \omega_f^2)^2 + 4\beta^2 \omega_f^2}} \quad (7.20)$$

$$tg(\Delta\phi) = \frac{\omega_f R/L}{\omega_0^2 - \omega_f^2} \quad (7.21)$$

The charge oscillating in the circuit is maximal when the angular frequency of the external driving is

$$\omega_Q = \sqrt{\omega_0^2 - 2\beta^2} \quad (7.22)$$

As the voltage of the capacitor is proportional to the charge stored in it, this is called voltage or charge resonance.

The current is:

$$I = \frac{dQ}{dt} = -Q_0\omega_f \sin(\omega_f t + \Delta\phi) \quad (7.23)$$

The amplitude of the current becomes maximal, when the frequency of the external driving force is equal to the natural frequency of the circuit. This is called current resonance.

7.2 Electromagnetic waves in perfect vacuum

As we have seen in the previous semester disturbances in an elastic medium may generate mechanical waves. In a similar manner disturbances of the electric and magnetic fields may generate electromagnetic waves. But unlike mechanical disturbances, electric and magnetic fields may exist in perfect vacuum. Therefore electromagnetic waves do not require the presence of any medium. Applying the Maxwell equations to perfect vacuum shows that changes in the electric field can generate a changing magnetic field, which in turn generates a changing electric field. From Faraday's law of induction:

$$\text{rot}\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \quad (7.24)$$

$$\text{rot}(\text{rot}\mathbf{E}) = -\text{rot}\frac{\partial\mathbf{B}}{\partial t} \quad (7.25)$$

$$\text{rot}(\text{rot}\mathbf{E}) = -\text{rot}\frac{\partial\mu_0\mathbf{H}}{\partial t} \quad (7.26)$$

$$\text{rot}(\text{rot}\mathbf{E}) = -\mu_0\frac{\partial}{\partial t}\text{rot}\mathbf{H} \quad (7.27)$$

In perfect vacuum there are no free charges, and consequently $\mathbf{j} = \mathbf{0}$. Thus the first term on the right-hand side of Ampere's law disappears:

$$\text{rot}\mathbf{H} = \mathbf{j} + \frac{\partial\mathbf{D}}{\partial t} = \frac{\partial\mathbf{D}}{\partial t} \quad (7.28)$$

Substituting this to (7.27) gives:

$$\text{rot}(\text{rot}\mathbf{E}) = -\mu_0\frac{\partial^2\mathbf{D}}{\partial t^2} \quad (7.29)$$

$$\text{rot}(\text{rot}\mathbf{E}) = -\mu_0\epsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2} \quad (7.30)$$

Using the mathematical identity $\text{rot}(\text{rot}\mathbf{E}) = \text{grad}(\text{div}\mathbf{E}) - \Delta\mathbf{E}$

$$\text{grad}(\text{div}\mathbf{E}) - \Delta\mathbf{E} = -\mu_0\epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \quad (7.31)$$

In perfect vacuum the charge density is 0, therefore:

$$\text{div}\mathbf{D} = \rho = 0 \quad (7.32)$$

$$\epsilon_0 \text{div}\mathbf{E} = 0 \quad (7.33)$$

Substituting this into (7.31), the first term on the left-hand side disappears:

$$\Delta\mathbf{E} = \mu_0\epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \quad (7.34)$$

This is a wave equation: the second differential of the electric field with respect to the space coordinate is proportional to the second differential with respect to time, respectively. The velocity of the wave may be determined by substituting a wave function into (7.34). (7.35) represents a plane wave:

$$\mathbf{E} = \mathbf{E}_0 e^{j(\omega t - \mathbf{k}\mathbf{r})} \quad (7.35)$$

Substituting this into (7.34) gives:

$$-\mathbf{k}\mathbf{k}\mathbf{E}_0 e^{j(\omega t - \mathbf{k}\mathbf{r})} = -\omega^2 \mu_0 \epsilon_0 \mathbf{E}_0 e^{j(\omega t - \mathbf{k}\mathbf{r})} \quad (7.36)$$

$$k^2 = \omega^2 \mu_0 \epsilon_0 \quad (7.37)$$

$$c = \frac{\omega}{k} = \frac{1}{\sqrt{\mu_0 \epsilon_0}} \quad (7.38)$$

Surprisingly this is exactly the same as the speed of light in vacuum. As scientists have already known that light exhibits wavelike behaviour, Maxwell and James Clerk have suggested in an 1865 paper that light may be a form of electromagnetic waves. Since then the hypothesis was proven by many experiments. Today we know that visible light along with radio waves, microwaves, X-rays and gamma radiation is indeed a form of electromagnetic waves. The lowest frequency (longest wavelength) electromagnetic waves are usually referred to as radio waves. Their wavelengths are typically several meters and their frequencies ranges up to a few hundred megahertz. Microwaves have frequencies in the gigahertz range and their wavelengths are usually a few centimetres or millimetres. Infrared radiation has a typical wavelength of a few microns, and its frequency is in the terahertz range. Visible light is a very narrow band in the electromagnetic spectrum: our eyes are capable of detecting electromagnetic waves whose wavelength is between 400 (violet) and 750 nm (red). If the wavelength is shorter than 400 nm, it becomes invisible to human eyes, and it is usually referred to as ultraviolet radiation. X-rays have a typical

wavelength of a few nanometres or only a few angstroms. Electromagnetic waves with wavelengths shorter than 10^{-2} nm are called gamma-rays.

Radio waves and microwaves are usually generated by accelerating charged particles (like the electrons in a radio antenna, or in the magnetron of a microwave oven). Microwaves and infrared radiation may also be created by molecular rotations. Most of the thermal radiation of room temperature bodies also appears in the infrared regime. Near-infrared radiation (the section of the infrared spectrum that is closest to visible light) and visible light may be the result of molecular vibrations. Visible light and ultraviolet radiation may also be created by electronic transitions in molecules and atoms. X-rays are usually generated by high energy electronic transitions in atoms, or by accelerating (or decelerating) high energy charged particles (bremsstrahlung). The highest frequency electromagnetic waves, the so called gamma-rays are related to nuclear processes.

7.3 Electromagnetic waves in non-conductive media

In the previous section we have assumed that both \mathbf{j} and ρ is zero. This is certainly true for perfect vacuum, but in the presence of a medium electric and magnetic fields may induce currents and polarise the medium, thus the deduction above loses validity. However in non-conductive media (such as glass, or different types of polymers) the assumption that the current density is zero still holds and polarisation can be taken into account through the dielectric constant and the permeability. In isotropic media:

$$\mathbf{D} = \epsilon_0 \epsilon_r \mathbf{E} = \epsilon \mathbf{E} \quad (7.39)$$

$$\mathbf{B} = \mu_0 \mu_r \mathbf{H} = \mu \mathbf{H} \quad (7.40)$$

Using these formulae, the deduction is very similar to what we have seen in the previous section. The only difference is that ϵ_0 and μ_0 is replaced by ϵ and μ . The wave equation becomes:

$$\Delta \mathbf{E} = \mu \epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} \quad (7.41)$$

And the speed of light in the medium is:

$$c = \frac{1}{\sqrt{\mu \epsilon}} \quad (7.42)$$

This speed is always lower than the speed of light in vacuum. The ratio of the speed of light in a medium to the speed of light in vacuum is called the index of refraction of the medium:

$$n = \frac{c_{medium}}{c_{vacuum}} = \frac{\frac{1}{\sqrt{\mu\epsilon}}}{\frac{1}{\sqrt{\mu_0\epsilon_0}}} = \frac{1}{\sqrt{\mu_r\epsilon_r}} \quad (7.43)$$

7.4 Direction of the \mathbf{E} and \mathbf{B} fields

The magnetic field can be determined by substituting (7.35) into Faraday's law of induction:

$$rot\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \quad (7.44)$$

$$\nabla \times \mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \quad (7.45)$$

$$j\mathbf{k} \times \mathbf{E}_0 e^{j(\omega t - \mathbf{k}\cdot\mathbf{r})} = -\frac{\partial\mathbf{B}}{\partial t} \quad (7.46)$$

$$\mathbf{B} = \frac{1}{\omega} \mathbf{k} \times \mathbf{E} \quad (7.47)$$

Therefore in free space the magnetic field is perpendicular to both the electric field and also to the wavenumber vector (\mathbf{k}). In a similar manner it can also be shown, that \mathbf{E} is also perpendicular to the \mathbf{B} and \mathbf{k} vectors. (It can be proven that this is true not only in free space, but also in any isotropic media.) As the wave is traveling in the direction of \mathbf{k} both the \mathbf{E} and \mathbf{B} fields are at right angles to the direction of propagation. In other words electromagnetic waves are transverse waves.

7.5 Pointing Vector

Waves transport energy and momentum. The respective energy densities of the electric and magnetic fields are:

$$u_E = \frac{1}{2}\epsilon_0 \mathbf{E}^2 \quad (7.48)$$

$$u_B = \frac{1}{2\mu_0} \mathbf{B}^2 \quad (7.49)$$

The amount of energy carried by the wave trough a given dA crossection in dt time is:

$$U = (u_E + u_B)dAcdt \quad (7.50)$$

The energy transfer through a unit of area in a unit of time is:

$$S = \frac{(u_E + u_B)dAcdt}{dAdt} \quad (7.51)$$

$$S = \left(\frac{1}{2}\epsilon_0 \mathbf{E}^2 + \frac{1}{2\mu_0} \mathbf{B}^2\right)c \quad (7.52)$$

According to (7.47) the intensity of the electric and magnetic fields are proportional to each other

$$|\mathbf{B}| = \frac{1}{\omega} |\mathbf{k} \times \mathbf{E}| \quad (7.53)$$

If \mathbf{k} and \mathbf{E} are perpendicular to each other

$$|\mathbf{B}| = \frac{|\mathbf{k}|}{\omega} |\mathbf{E}| \quad (7.54)$$

$$|\mathbf{B}| = \frac{1}{c} |\mathbf{E}| \quad (7.55)$$

Substituting this into (7.52)

$$S = \frac{1}{2}(\epsilon_0 c |\mathbf{E}| |\mathbf{B}| + \frac{1}{\mu_0 c} |\mathbf{E}| |\mathbf{B}|)c \quad (7.56)$$

$$S = \frac{|\mathbf{E}| |\mathbf{B}|}{2} (\epsilon_0 c^2 + \frac{1}{\mu_0}) \quad (7.57)$$

$$S = \frac{|\mathbf{E}| |\mathbf{B}|}{\mu_0} \quad (7.58)$$

It is known that the wave is traveling in the direction of the wavenumber vector, and this vector is perpendicular to both \mathbf{E} and \mathbf{B} . Therefore energy is transported in the direction of $\mathbf{E} \times \mathbf{B}$. With this knowledge we may introduce the Poynting vector, whose magnitude gives the amount of energy transported through a unit of area in a unit of time, and points in the direction of the energy transport:

$$\mathbf{S} = \frac{1}{\mu_0} \mathbf{E} \times \mathbf{B} \quad (7.59)$$

7.6 Light-pressure

Besides energy, a wave may also transport momentum. When it is reflected back from a surface its momentum changes direction. As momentum should be conserved this is only possible, if there is a momentum transfer between the wave and the surface. In other words the wave exerts a force on the surface from which it is reflected. By calculating this force we may determine the momentum carried by the wave.

So far we have considered electromagnetic waves traveling in perfect vacuum and in non-conductive media. But electric and magnetic fields can interact only with charged particles, such as electrons in a metal. It is through these charged particles that the wave can exert a force on the surface: it may influence the movement of the charged particles, then the particles may transfer the momentum received from the wave to the rest of the medium. Therefore to calculate the force we have to describe the propagation of electromagnetic waves in conductive media.

Let us consider a wave that arrives perpendicularly to a metal surface. We already know that the electric and magnetic fields of the wave are perpendicular to each other, and also to the direction of propagation. Let us choose the axis of our coordinate system so that the wave is traveling in the positive x direction, \mathbf{E} is parallel to the y axis, and \mathbf{B} is pointing in the direction of the z axis. In this case the intensities of the electric and magnetic fields on the surface are:

$$\mathbf{E} = (E_0 \sin(\omega t)) \mathbf{u}_y \quad (7.60)$$

$$\mathbf{B} = (B_0 \sin(\omega t)) \mathbf{u}_z \quad (7.61)$$

where \mathbf{u}_y and \mathbf{u}_z are unit vectors pointing in the direction of the y and z axes, respectively. The force exerted on charged particles by the electric field is parallel to \mathbf{E} and forces them to start oscillating in the y direction. But the amplitude of these oscillations and the velocity of the particles cannot increase to infinity due to collisions with other particles in the solid. These losses act as if the charged particles are moving in a viscous medium. The velocity increases until the drag force becomes equal to the force exerted on the particles by the electric field.

$$\mathbf{F}_E = (qE_0 \sin(\omega t)) \mathbf{u}_y \quad (7.62)$$

$$\mathbf{F}_D = k \mathbf{v}_y \quad (7.63)$$

where q is the charge of the particles, and k is a constant representing the “viscosity” of the medium in which the particles are moving. Comparing the two equations gives:

$$\mathbf{v}_y = \left(\frac{qE_0}{k} \sin(\omega t) \right) \mathbf{u}_y \quad (7.64)$$

The magnetic field can also exert a force on moving charged particles:

$$\mathbf{F}_B = q \mathbf{v}_y \times \mathbf{B} \quad (7.65)$$

$$\mathbf{F}_B = q \left(\frac{qE_0}{k} \sin(\omega t) \right) \mathbf{u}_y \times (B_0 \sin(\omega t)) \mathbf{u}_z \quad (7.66)$$

$$\mathbf{F}_B = \left(\frac{q^2 E_0 B_0}{k} \sin^2(\omega t) \right) \mathbf{u}_x \quad (7.67)$$

Unlike the force exerted on charged particles by the electric field, which changes sign twice every cycle, and averages out in a longer period of time, this force always points

toward the positive x direction. When averaged over a longer period of time (compared to the period), it results in a net force perpendicular to the surface. This is the force representing the momentum transfer between the electromagnetic wave and the surface. The energy transferred to a particle in a unit of time is:

$$\frac{dU}{dt} = (\mathbf{F}_E + \mathbf{F}_B) \cdot \mathbf{v} \quad (7.68)$$

As the force exerted on the particles by the magnetic field is perpendicular to their velocity it can be ignored:

$$\frac{dU}{dt} = (qE_0 \sin(\omega t)) \mathbf{u}_y \cdot \left(\frac{qE_0}{k} \sin(\omega t) \right) \mathbf{u}_y \quad (7.69)$$

$$\frac{dU}{dt} = \frac{q^2 E_0^2}{k} \sin^2(\omega t) = c \frac{q^2 E_0 B_0}{k} \sin^2(\omega t) \quad (7.70)$$

Comparing equations (7.67) and (7.70) shows that the power transferred to the particle is proportional to the force exerted on it by the magnetic field:

$$|\mathbf{F}_B| = \frac{1}{c} \frac{dU}{dt} \quad (7.71)$$

(7.71) show that the momentum transfer in a unit of time on a unit of surface area is proportional to the energy transfer in the same amount of time on the same area. We already know that the amount of energy carried by the electromagnetic wave is given by the Pointing vector. This means, that the momentum of the wave is proportional to the Pointing vector:

$$\frac{d\mathbf{p}}{dA dt} = \frac{\mathbf{S}}{c} \quad (7.72)$$

The light pressure is the force with which the light acts on a unit area of the surface. When light is adsorbed, it transfers all of its momentum to the surface. In this case the light-pressure is:

$$P_{abs} = \frac{dF}{dA} = \frac{dp}{dA dt} = \frac{|\mathbf{S}|}{c} \quad (7.73)$$

When light is reflected back from a surface, its momentum changes sign. This means, that the momentum transfer between the electromagnetic wave and the surface is twice as intense as in case of absorption.

$$P_{refl} = 2 \frac{|\mathbf{S}|}{c} \quad (7.74)$$

7.7 Skin depth

In section 7.2 we have deduced the existence of electromagnetic waves in perfect vacuum. But as we have seen in the previous section electromagnetic waves can transfer part of their energy and momentum to charged particles. Due to these losses the intensity of the waves decreases in conductive media. To take this into account the simplified model of section 7.2 needs to be augmented:

$$\text{rot}\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \quad (7.75)$$

$$\text{rot}(\text{rot}\mathbf{E}) = -\text{rot}\frac{\partial\mathbf{B}}{\partial t} \quad (7.76)$$

$$\text{rot}(\text{rot}\mathbf{E}) = -\text{rot}\frac{\partial\mu\mathbf{H}}{\partial t} \quad (7.77)$$

$$\text{rot}(\text{rot}\mathbf{E}) = -\mu\frac{\partial}{\partial t}\text{rot}\mathbf{H} \quad (7.78)$$

According to the third Maxwell equation:

$$\text{rot}\mathbf{H} = \mathbf{j} + \frac{\partial\mathbf{D}}{\partial t} \quad (7.79)$$

In perfect vacuum, the first term on the left-hand side could be ignored. But in conductive media the electric field forces charged particles to move, therefore $\mathbf{j} = \sigma\mathbf{E}$ (where σ is the conductivity of the medium)

$$\text{rot}\mathbf{H} = \sigma\mathbf{E} + \epsilon\frac{\partial\mathbf{E}}{\partial t} \quad (7.80)$$

Substituting this into (7.78):

$$\text{rot}(\text{rot}\mathbf{E}) = -\mu\epsilon\frac{\partial^2\mathbf{E}}{\partial t^2} - \mu\sigma\frac{\partial\mathbf{E}}{\partial t} \quad (7.81)$$

Using the mathematical identity $\text{rot}(\text{rot}\mathbf{E}) = \text{grad}(\text{div}\mathbf{E}) - \Delta\mathbf{E}$

$$\text{grad}(\text{div}\mathbf{E}) - \Delta\mathbf{E} = -\mu\epsilon\frac{\partial^2\mathbf{E}}{\partial t^2} - \mu\sigma\frac{\partial\mathbf{E}}{\partial t} \quad (7.82)$$

If there is no space-charge:

$$\text{div}\mathbf{D} = \rho = 0 \quad (7.83)$$

$$\epsilon\text{div}\mathbf{E} = 0 \quad (7.84)$$

Substituting this into (7.82), the first term on the left-hand side disappears:

$$\Delta\mathbf{E} = \mu\epsilon\frac{\partial^2\mathbf{E}}{\partial t^2} + \mu\sigma\frac{\partial\mathbf{E}}{\partial t} \quad (7.85)$$

The equation is very similar to (7.34), except for the last term, which represent losses due to the energy transferred to the conductive medium. Substituting the wave function of a plane wave into (7.85) as a trial function gives:

$$\mathbf{E} = \mathbf{E}_0 e^{j(\omega t - \mathbf{k}\mathbf{r})} \quad (7.86)$$

$$-\mathbf{k}\mathbf{k}\mathbf{E}_0 e^{j(\omega t - \mathbf{k}\mathbf{r})} = -\omega^2 \mu \epsilon \mathbf{E}_0 e^{j(\omega t - \mathbf{k}\mathbf{r})} \quad (7.87)$$

$$\mathbf{k}^2 = \mu \epsilon \omega^2 + j \sigma \mu \omega \quad (7.88)$$

$$|\mathbf{k}| = \sqrt{\mu \epsilon \omega^2 + j \sigma \mu \omega} \quad (7.89)$$

For metals, the first term on the right-hand side is negligible compared to the second term:

$$|\mathbf{k}| = \sqrt{j \sigma \mu \omega} \quad (7.90)$$

$$|\mathbf{k}| = \sqrt{\frac{\sigma \mu \omega}{2}} (1 + j) \quad (7.91)$$

$$|\mathbf{k}| = k_r + j k_{im} \quad (7.92)$$

where

$$k_r = k_{im} = \sqrt{\frac{\sigma \mu \omega}{2}} \quad (7.93)$$

Therefore \mathbf{k} have an imaginary component. (The situation is similar to the underdamped case of the damped oscillator, where the imaginary component of the angular frequency caused an exponential decay of the amplitude.) Substituting (7.93) back into (7.35) gives:

$$\mathbf{E} = \mathbf{E}_0 e^{j(\omega t - \mathbf{k}\mathbf{r})} \quad (7.94)$$

$$\mathbf{E} = \mathbf{E}_0 e^{j(\omega t - (\mathbf{k}_r + j\mathbf{k}_{im})\mathbf{r})} \quad (7.95)$$

$$\mathbf{E} = \mathbf{E}_0 e^{-\mathbf{k}_{im}\mathbf{r}} e^{j(\omega t - \mathbf{k}_r\mathbf{r})} \quad (7.96)$$

Therefore the amplitude of electromagnetic waves decay exponentially in conductive media. Due to the $e^{-\mathbf{k}_{im}\mathbf{r}}$ term the rate of this decay depends on $k_{im} = \sqrt{\frac{\sigma \mu \omega}{2}}$. The distance in which the amplitude of the wave decreases by a factor of e is called skin depth:

$$\delta = \frac{1}{k_{im}} = \sqrt{\frac{2}{\sigma \mu \omega}} \quad (7.97)$$

This result helps us to choose suitable materials for electromagnetic shielding. Skin depth depends not only on the conductivity of the medium, but also on its magnetic

permeability. Although copper and aluminium are good conductors, their permeabilities are low, therefore (contrary to common belief) they are not ideal for electromagnetic shielding. Several different alloys (such as permalloy and mu metal) have been developed for this purpose. Although they are not as good conductors as copper, their permeability is several orders of magnitude higher, which makes them much more suitable for shielding.

Another interesting property of skin depth is that it depends not only on the properties of the medium, but also on the frequency of the electromagnetic wave. High frequency waves transfer their energy to the medium very quickly, and they have very small skin depth. As the frequency of visible light is very high, its skin depth is very small. This is the reason why metals are not transparent. At lower frequencies electromagnetic waves may penetrate deep into the medium. This also means that shielding against low frequency electromagnetic interference is considerably harder.

This is also the reason why nuclear submarines use very low frequency radio waves to communicate with their command centre: seawater is a conductive medium¹, and it adsorbs radio waves. By using lower frequencies skin depth may be increased and the submarine may stay in contact with its command centre even while it is submerged. (It must be noted however that low frequency communication has a very limited bandwidth, therefore such communication channels are used only to transfer the most crucial commands...)

7.8 Reflection and refraction

In the previous section it was concluded that visible light cannot penetrate deep into conductive media. But we know from practical experience that there are transparent materials such as glass and different types of polymers. (A common characteristic of these materials is that they are non-conductive.) In this section we shall discuss the behaviour of electromagnetic radiation at the interface between two such non-conductive materials.

Imagine an electromagnetic wave arriving to the boundary. Part of it may be reflected back from the interface, the rest will be transmitted. Therefore we shall consider three electromagnetic waves: the incident wave (whose electric field is \mathbf{E}_i), the reflected wave

¹It must be noted, that although water is a conductive medium, it is transparent for visible light. The reason of this is that in water it is not electrons that conduct electric currents but the ions of different kinds of salts. These ions are considerably less mobile than the electrons in metals. Therefore they are unable to follow the quick changes of high frequency electromagnetic waves. Although water is perfectly capable to adsorb radio waves the above described mechanism does not work at higher frequencies. (It must be noted however that other regimes of the electromagnetic spectrum may be adsorbed by other mechanisms.)

(\mathbf{E}_r) and the transmitted wave (\mathbf{E}_t):

$$\mathbf{E}_i = \mathbf{E}_{i0} e^{j(\mathbf{k}_i \mathbf{r} - \omega_i t)} \quad (7.98)$$

$$\mathbf{E}_r = \mathbf{E}_{r0} e^{j(\mathbf{k}_r \mathbf{r} - \omega_r t)} \quad (7.99)$$

$$\mathbf{E}_t = \mathbf{E}_{t0} e^{j(\mathbf{k}_t \mathbf{r} - \omega_t t)} \quad (7.100)$$

At the boundary between the two media all three waves must exist simultaneously and the tangential components shall be equal on both sides. In mathematical terms:

$$\mathbf{n} \times \mathbf{E}_i + \mathbf{n} \times \mathbf{E}_r = \mathbf{n} \times \mathbf{E}_t \quad (7.101)$$

Where \mathbf{n} is the normal vector of the interface.

$$\mathbf{n} \times \mathbf{E}_{i0} e^{j(\mathbf{k}_i \mathbf{r} - \omega_i t)} + \mathbf{n} \times \mathbf{E}_{r0} e^{j(\mathbf{k}_r \mathbf{r} - \omega_r t)} = \mathbf{n} \times \mathbf{E}_{t0} e^{j(\mathbf{k}_t \mathbf{r} - \omega_t t)} \quad (7.102)$$

The equation above holds for all moments of time at all points of the interface only if:

$$\mathbf{k}_i \mathbf{r} - \omega_i t = \mathbf{k}_r \mathbf{r} - \omega_r t = \mathbf{k}_t \mathbf{r} - \omega_t t \quad (7.103)$$

This means that the phases of all three waves should be the same at the interface at all moments of time. This also means that it should hold for $\mathbf{r} = \mathbf{0}$:

$$\omega_i t = \omega_r t = \omega_t t \quad (7.104)$$

$$\omega_i = \omega_r = \omega_t \quad (7.105)$$

Therefore the frequencies of all three waves are the same. Reflection and refraction may change the direction of propagation (and in case of refraction even the wavelength may change), but the frequency always remains the same.

We may also use (7.103) to determine the direction into which the wave is reflected or refracted. As the equation is true for all moments of time, it should also hold for $t = 0$:

$$\mathbf{k}_i \mathbf{r} = \mathbf{k}_r \mathbf{r} = \mathbf{k}_t \mathbf{r} \quad (7.106)$$

From $\mathbf{k}_i \mathbf{r} = \mathbf{k}_r \mathbf{r}$:

$$k_i r \sin \theta_i = k_r r \sin \theta_r \quad (7.107)$$

Both the incident and the reflected wave travels in the same medium, therefore their velocities are identical. Since their frequency is also the same, $k_i = k_r$. Therefore:

$$\sin \theta_i = \sin \theta_r \quad (7.108)$$

$$\theta_i = \theta_r \quad (7.109)$$

This means, that reflection is symmetrical: the angle between the incident wave and the normal of the interface is the same as the angle between the reflected wave and the normal. This is called the law of reflection. The principle is utilised in a wide range of optical systems from telescopes to the headlights of cars.

Form $\mathbf{k}_i \mathbf{r} = \mathbf{k}_t \mathbf{r}$:

$$k_i r \sin \theta_i = k_t r \sin \theta_t \quad (7.110)$$

$$\frac{\sin \theta_t}{\sin \theta_i} = \frac{k_i}{k_t} = \frac{v_i}{v_t} = \frac{n_1 c}{n_2 c} \quad (7.111)$$

$$\frac{\sin \theta_t}{\sin \theta_i} = \frac{n_1}{n_2} \quad (7.112)$$

where n_1 is the index of refraction of the first medium (from which the wave arrives to the interface) and n_2 is the index of refraction of the second medium (in which the transmitted wave travels). (7.112) is called the law of refraction or Snell's law. It shows that electromagnetic waves cannot travel through the interface without changing their direction of propagation. This phenomenon has widespread applications in optics. Lenses, optical fibres, and prisms are all based on this phenomenon. Even rainbows appear due to the fact that the index of refraction depends on the wavelength of light, therefore different colours are refracted into different directions by the tiny drops of water floating in air after a rain.

Chapter 8

Geometrical Optics - Gábor Dobos

In the previous chapter we have seen that light is basically a type of electromagnetic waves. But in many cases its behaviour can be described by a much simpler model. When the characteristic size of the structures with which light interacts is considerably larger than its wavelength, it travels in a straight line in homogenous media, and changes its direction only when the medium's index of refraction changes. (At the interface of two different materials it may be reflected or refracted.) In these cases we don't have to solve the Maxwell equations to calculate the wave function: the behaviour of light can be described by simple geometrical methods.

In these cases we may imagine light as rays, traveling in straight lines and use the laws of reflection and refraction to describe how its trajectory changes at the surface of different optical elements. Since the wavelength of visible light is only a few hundred nanometres and the characteristic sizes of lenses and mirrors are typically in the centimetre or metre range, in most cases no wavelike behaviour can be observed. (In fact the laws of geometrical optics were discovered before scientists even realised that light is a wave.)

In this chapter we will discuss classical optical systems using geometrical optics. It must be noted however that this is merely an approximation, applicable only in the aforementioned cases. Those special cases where one dimension of the optical system becomes comparable to the wavelength of light, and the simplified model of geometrical optics is no longer applicable, will be discussed in the next chapter.

8.1 Total internal reflection

In the previous chapter we have deduced the law of refraction, also known as Snell's law:

$$\frac{\sin\theta_r}{\sin\theta_i} = \frac{n_1}{n_2} \quad (8.1)$$

Where the so called incident angle (θ_i) is the angle between the incident ray of light and the normal of the surface, and the refraction angle (θ_r) is the angle between the refracted ray of light and the normal. The refractive index of the medium from which the incident ray of light arrives to the interface is n_1 , and n_2 is the refractive index of the other medium. The equation shows that when light is traveling from an optically denser material to an optically rarer material ($n_1 > n_2$) the refraction angle is higher than the incident angle. This means that there is a critical incident angle (which is smaller than 90 degrees), at which the refraction angle reaches 90 degrees. Since in this case light cannot pass through the interface, it is completely reflected back to the optically denser medium following the law of reflection. This is called total internal reflection.

The phenomenon has several practical applications. Many optical systems apply prisms instead of mirrors. A prism is basically a piece of glass with flat polished surfaces. Light may freely enter the prism through a side that is arranged to be perpendicular to its trajectory. The prism is cut in such a way that the ray of light reaches the second side of the prism at an angle that is higher than the aforementioned critical angle thus it undergoes total internal reflection, and continues its trajectory as if it was reflected on a mirror. A third side of the prism is cut in an angle that is perpendicular to the new direction of the ray, so that it can leave the prism.

A similar principle can be used to guide light to large distances. For example a simple optical fibre consists of a transparent core, surrounded by a cladding material with a lower index of refraction. The end of the fibre is usually perpendicular to its axis, so light can enter the core easily from the axial direction. These rays of light will reach the interface between the core and the cladding at a flat angle. Due to total internal reflection they bounce back from the interface and continue their trajectory inside the core. Since reflection is symmetrical, the new direction is also close to the axial direction, and the ray reaches the interface at flat angles over and over again and bounces back to the core, without ever leaving it. This way light can be guided to large distances inside to core of the fibre.

8.2 Spherical Mirror

From the headlights of cars to astronomical telescopes many different optical systems contain curved mirrors. Their exact shapes may vary depending on the application, but the simplest type is the so-called spherical mirror. It is relatively easy to describe its behaviour in mathematical terms, and the principles we may deduce for it may be applicable to other types of curved mirrors, too.

To describe spherical mirrors, we have to consider three distinct cases. The first one is depicted in figure 8.1. Point C marks the centre of the sphere, while M is the centre of the mirror. The line drawn through these points is called the optical axis. Imagine that an object on the optical axis (in point O) emits a ray of light, which is reflected back

from the mirror in point P . Let us calculate the position where this reflected ray will cross the optical axis (point I)!

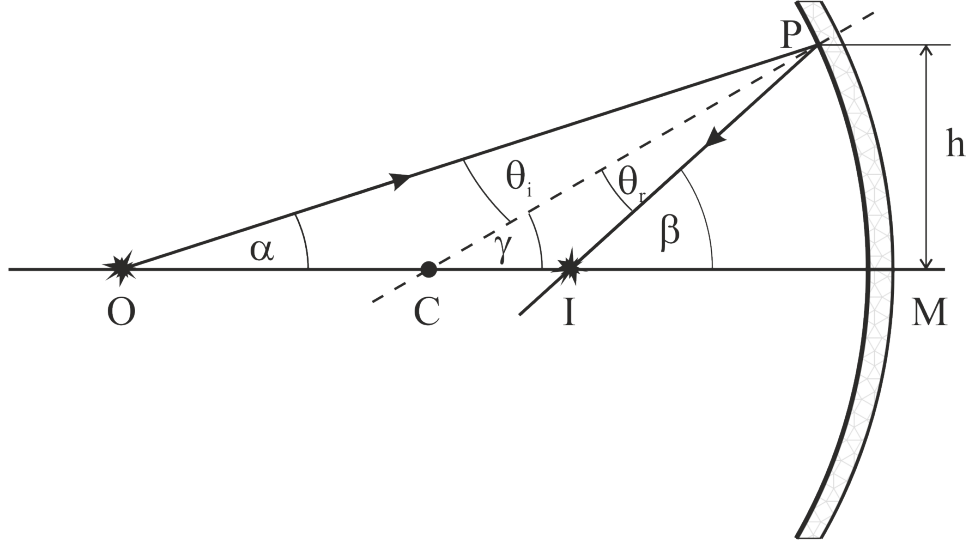


Figure 8.1: Reflection on a concave mirror if the light source is farther away from the mirror than its radius

The CP segment is the radius of the sphere thus it is perpendicular to the surface of the mirror. Therefore the incident and reflected ray should be symmetrical to this line. In other words, the angle between the incident ray of light and the dashed line (the incident angle) is the same as the angle between the dashed line and the reflected ray of light (reflection angle):

$$\theta_i = \theta_r = \theta \quad (8.2)$$

The β angle is the exterior angle of both the OIP and CIP triangles. Since the exterior angle of a triangle is the sum of the two remote interior angles we may calculate β from the angles of both triangles:

$$\beta = \alpha + 2\theta \quad (8.3)$$

$$\beta = \gamma + \theta \quad (8.4)$$

We may eliminate θ by combining these two equations:

$$\alpha + \beta = 2\gamma \quad (8.5)$$

The α angle may be determined from the MOP triangle. If it is small enough, point P is so close to point M , that the curvature of the mirror becomes negligible and the

$OMP\angle$ angle is close to 90 degrees. In this case $\sin\alpha = \frac{h}{OM}$. But for small angles $\sin\alpha$ may be approximated by α , therefore $\alpha \approx \frac{h}{OM}$. Since this approximation is applicable only when the rays of light are almost parallel to the optical axis, it is usually referred to as paraxial approximation.

The γ and β angles may be estimated in a similar manner. From the PCM triangle $\gamma \approx \frac{h}{CM}$, and from the PIM triangle $\beta \approx \frac{h}{IM}$. Substituting these into (8.5) gives:

$$\frac{h}{OM} + \frac{h}{IM} = 2\frac{h}{CM} \quad (8.6)$$

$$\frac{1}{OM} + \frac{1}{IM} = \frac{2}{CM} \quad (8.7)$$

An interesting feature of this equation is that it is independent of the angles between the rays and the optical axis (as long as they are suitably small). All rays of light emitted by the object (in a suitably small angle) will be focused to the same point: all reflected rays will cross each other in point I , and continue their trajectories as if they have originated from that point.

When we look at objects in the physical world, our eyes detect the light scattered on their surface. Basically all points of the object act like a light source. Therefore if we place an object in point O the mirror will project the rays of light scattered on its surface to point I , and they will continue their trajectory as if they have originated from that point. An observer looking into the optical system will perceive these reflected rays, as if they have been scattered on the surface of an object in point I . In other words the concave spherical mirror projects an image of the object to point I .

It must be noted, that the situation is slightly different if the object is closer to the mirror than its radius. This second case is depicted in figure 8.2. Again, the CP segment is the radius of the sphere, thus

$$\theta_i = \theta_r = \theta \quad (8.8)$$

Again, the β angle may be determined from both the OIP and CIP triangles. (But in this case β is one of the internal angles of both triangles. The external angles are $\theta_i + \theta_r = 2\theta$ and $\theta_r = \theta$, respectively.)

$$2\theta = \alpha + \beta \quad (8.9)$$

$$\theta = \gamma + \beta \quad (8.10)$$

θ may be eliminated by combining the two equations:

$$\alpha - \beta = 2\gamma \quad (8.11)$$

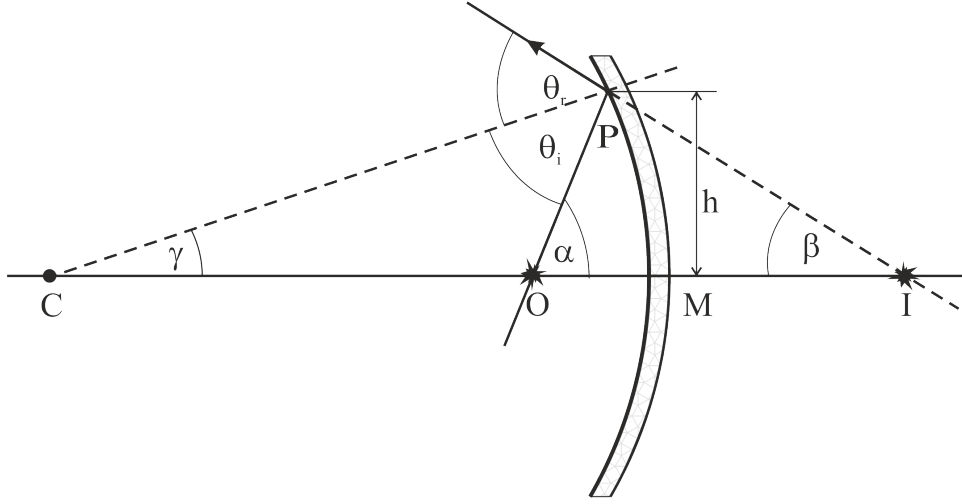


Figure 8.2: Reflection on a concave mirror if the light source is closer to the mirror than its radius

Using the paraxial approximation:

$$\frac{1}{OM} - \frac{1}{IM} = \frac{2}{CM} \quad (8.12)$$

The third case to consider is when the object is placed in front of a convex mirror (figure 8.3). From the OPI and OPC triangles:

$$2\theta = \alpha + \beta \quad (8.13)$$

$$\theta = \alpha + \gamma \quad (8.14)$$

Combining the two equations to eliminate θ gives:

$$\alpha - \beta = -2\gamma \quad (8.15)$$

Using the paraxial approximation:

$$\frac{1}{OM} - \frac{1}{IM} = -\frac{2}{CM} \quad (8.16)$$

Note, that in the first case the image appears in front of the mirror, while in the second and third case it is behind it. Consequently in equation (8.7) the second term $\left(\frac{1}{IM}\right)$ has a positive sign while in equation (8.12) and (8.16) it has a negative sign. Also in the first two cases (where we have discussed concave mirrors) the right-hand side of the equation had a positive sign, while in the third case (where we have discussed the behaviour of

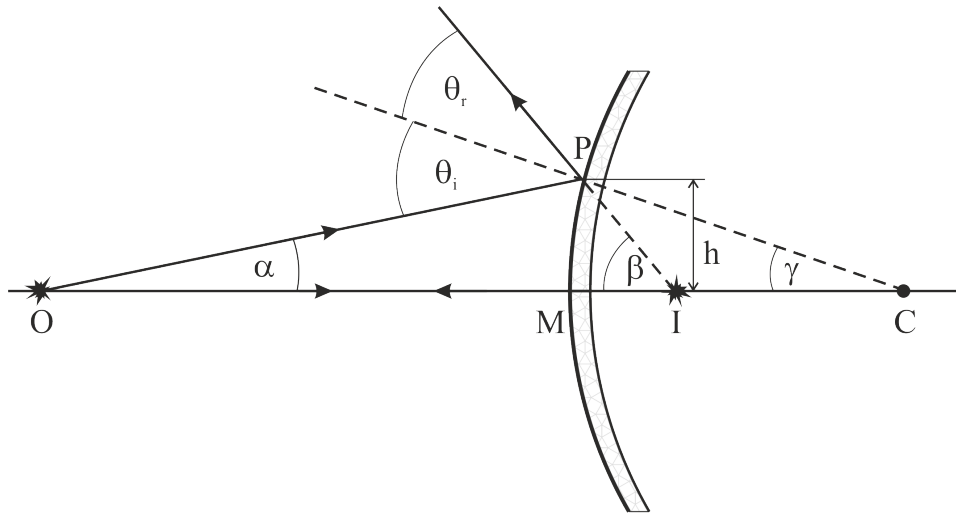


Figure 8.3: Reflection on a convex mirror

a convex mirror) the right-hand side had a negative sign. Except of these, the three equations are identical. Therefore we may express all three cases by a single equation.

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f} \quad (8.17)$$

- The parameter o is called object distance. It is the distance of the object (point O) from the centre of the mirror (point M). It has a positive sign if the object is in front of the mirror, and a negative sign if it is behind it. (The latter may occur when the mirror is part of a complex optical system, and the previous stage of the system projects the image of an object behind the mirror. In this case this image - which appears behind the mirror - serves as the object of the next projection.)
- The parameter i is called image distance. It is the distance of the image (point I) from the centre of the mirror (point M). It has a positive sign if the image is in front of the mirror, and a negative sign if it is behind it.
- The parameter f is called the focal distance of the mirror, because rays of light arriving to the mirror parallel to the optical axis will be focused to a point on the optical axis (the so called focal point) which is precisely f distance away from the centre of the mirror (point M). (This can be imagined as if light is coming from an object infinitely far away so that the divergence of the rays is negligible. In this case the object distance is infinite, and its inverse is zero. Therefore the image distance becomes equal to the focal distance.) For spherical mirrors f is half of the radius of curvature. The sign of the focal distance is positive for concave mirrors, and negative for convex mirrors.

Equation (8.17) is called the projection law, and it is applicable not only for spherical, but for other types of mirrors, too. Although their shapes are different, in the paraxial approximation parabolic and hyperbolic mirrors behave in a similar manner.

8.3 Thin spherical lenses

Optical systems may contain lenses instead of curved mirrors. These are usually made of glass, and just like spherical or parabolic mirrors, they may be used to focus light, and project images of objects. Whereas in case of spherical mirrors light was focused by reflection on a curved surface, in case of lenses light is refracted on the curved boundaries between air and glass. To describe the behaviour of lenses let us calculate how the direction of a ray of light changes when it enters the lens, and when it leaves it. For the sake of simplicity let us consider spherical surfaces yet again:

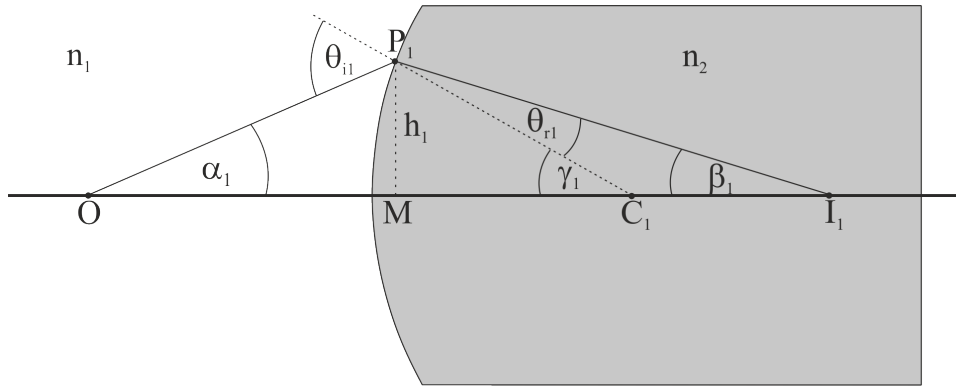


Figure 8.4: Refraction of light on the input surface of a spherical lens

On figure 8.4 a ray of light originating from point O enters the lens in point P_1 . Point C_1 marks the centre of the spherical surface, thus the P_1C_1 section is a radius of the sphere and it is perpendicular to the surface. The incident angle (θ_{i1}) is the angle between the OP_1 and P_1C_1 lines, whereas the refraction angle (θ_{r1}) is the angle between the perpendicular (P_1C_1 line) and the refracted ray of light (P_1I_1 line). The refractive indices of the outside medium and the glass are marked by n_1 and n_2 , respectively.

The θ_{i1} angle is the external angle of the OP_1C_1 triangle, therefore:

$$\theta_{i1} = \alpha_1 + \gamma_1 \quad (8.18)$$

The γ_1 angle is the external angle of the $P_1C_1I_1$ triangle, thus:

$$\gamma_1 = \theta_{r1} + \beta_1 \quad (8.19)$$

According to the law of refraction:

$$n_1 \sin(\theta_{i1}) = n_2 \sin(\theta_{r1}) \quad (8.20)$$

In the paraxial approximation:

$$n_1 \theta_{i1} = n_2 \theta_{r1} \quad (8.21)$$

θ_{i1} and θ_{r1} may be expressed from equations (8.18) and (8.19). Substituting these into equation (8.21) gives:

$$n_1 \alpha_1 + n_1 \gamma_1 = n_2 \gamma_1 - n_2 \beta_1 \quad (8.22)$$

$$n_1 \alpha_1 + n_2 \beta_1 = (n_2 - n_1) \gamma_1 \quad (8.23)$$

The angles α_1 , β_1 and γ_1 may be estimated in the paraxial approximation in a similar fashion as in the previous section:

$$\alpha_1 \approx \frac{h}{OM} \quad (8.24)$$

$$\beta_1 \approx \frac{h}{I_1 M} \quad (8.25)$$

$$\gamma_1 \approx \frac{h}{C_1 M} \quad (8.26)$$

Substituting these into equation (8.23) gives:

$$\frac{n_1}{OM} + \frac{n_2}{I_1 M} = \frac{n_2 - n_1}{C_1 M} \quad (8.27)$$

Or:

$$\frac{n_1}{o} + \frac{n_2}{i_1} = \frac{n_2 - n_1}{R_1} \quad (8.28)$$

where o is the object distance, R_1 is the radius of the input surface of the lens, and i_1 is the distance of point I_1 (where the ray crosses the optical axis) from the centre of the lens (point M). The only problem is, that actual lenses are usually thin, and therefore the ray leaves the lens before it could reach point I_1 . During this, it undergoes a second refraction, that can be described in a similar fashion as the first.

On figure 8.5 the ray of light arrives to point P_2 where it undergoes its second refraction. Whereas it would have crossed the optical axis in point I_1 , after the second refraction it will head towards point I . Point C_2 marks the centre of the spherical exit surface whose radius is R_2 . As the $C_2 P_2$ segment is the radius of the sphere it is perpendicular to the surface. The incident angle (the angle between the incoming ray of

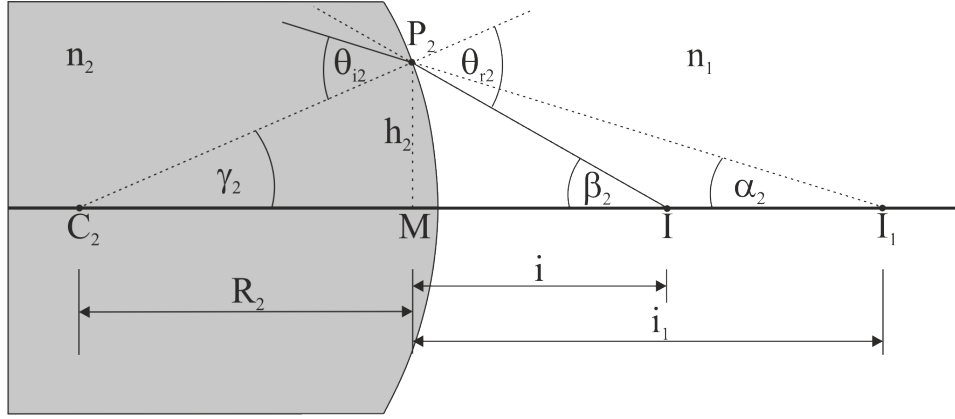


Figure 8.5: Refraction of light on the exit surface of a spherical lens

light and the perpendicular) is marked by θ_{i2} , while the refraction angle is marked by θ_{r2} . These angles are the external angles of the $C_2P_2I_1$ and C_2P_2I triangles, therefore:

$$\theta_{i2} = \gamma_2 + \alpha_2 \quad (8.29)$$

$$\theta_{r2} = \gamma_2 + \beta_2 \quad (8.30)$$

According to the law of refraction:

$$n_2 \sin(\theta_{i2}) = n_1 \sin(\theta_{r2}) \quad (8.31)$$

In the paraxial approximation:

$$n_2 \theta_{i2} = n_1 \theta_{r2} \quad (8.32)$$

Substituting θ_{i2} and θ_{r2} from (8.29) and (8.30) into this equation gives:

$$n_2 \gamma_2 + n_2 \alpha_2 = n_1 \gamma_2 + n_1 \beta_2 \quad (8.33)$$

$$n_1 \beta_2 - n_2 \alpha_2 = (n_2 - n_1) \gamma_2 \quad (8.34)$$

The α_2 , β_2 and γ_2 angles may be estimated from the P_2I_1M , P_2IM and P_2C_2M triangles, respectively:

$$\frac{n_1}{IM} - \frac{n_2}{I_1M} = \frac{n_2 - n_1}{C_2M} \quad (8.35)$$

Or:

$$\frac{n_1}{i} - \frac{n_2}{i_1} = \frac{n_2 - n_1}{R_2} \quad (8.36)$$

By comparing equations (8.28) and (8.36), the i_1 distance may be eliminated:

$$\frac{n_1}{o} + \frac{n_1}{i} = (n_2 - n_1) \left(\frac{1}{R_1} + \frac{1}{R_2} \right) \quad (8.37)$$

$$\frac{n_1}{o} + \frac{n_1}{i} = (n - 1) \left(\frac{1}{R_1} + \frac{1}{R_2} \right) \quad (8.38)$$

where $n = \frac{n_2}{n_1}$. Comparing this equation to the projection law for spherical mirrors, it is obvious, that the two are very similar to each other, and the focal distance of the lens is determined by the following formula:

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} + \frac{1}{R_2} \right) \quad (8.39)$$

This equation is commonly referred to as the lensmaker's equation. (Using this formula a lensmaker may estimate the necessary curvatures to form a lens of a given focal length.) The signs of the parameters in the equation depend on whether the surfaces are convex or concave. In this form of the equation both R_1 and R_2 are positive for convex surfaces and negative for concave surfaces. (The deduction for concave surfaces would be very similar to the one presented above, only the signs of certain terms would be different...)

It must be noted that we have made several approximations during the above deduction, and simplified the problem considerably. This means that the equation in this form is valid only for thin lenses (where the curvature radii are considerably larger than the thickness of the lens). For thick lenses the formula must be augmented by an additional term:

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} + \frac{1}{R_2} - \frac{(n - 1)d}{nR_1R_2} \right) \quad (8.40)$$

where d is the thickness of the lens.

8.4 Projection by spherical lenses and mirrors

As we have seen in the previous sections the behaviour of spherical lenses and mirrors are very similar. Both can be described by the projections law:

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f} \quad (8.41)$$

In general we may say that:

- The object distance (o) is the distance of the object from the centre of the lens or mirror, and it has a positive sign if the rays of light heading towards the lens or mirror are divergent, and a negative sign if they are convergent.

- The image distance (i) is the distance of the image from the centre of the lens or mirror, and it has a positive sign if the rays leaving the lens or mirror are convergent, and a negative sign if they are divergent.
- The focal distance (f) depends on the curvature radii. For spherical mirrors it is half of the radius of curvature, with a positive sign for concave mirrors, and a negative sign for convex mirrors. In case of lenses, the focal distance may be determined by the lensmaker's equation. The inverse of the focal distance measured in metres is usually called dioptré, and it is commonly used by opticians to define the optical power of prescription glasses.

Using the projection law we may calculate where the image of an object will appear. Alternatively we may use simple geometrical principles to follow certain characteristic rays of light originating from the object, and check where they cross each other to construct the image. These characteristic rays are:

- A ray that arrives parallel to the optical axis will be reflected back from the mirror or refracted by the lens so that it goes through the focal point.
- A ray arriving through the focal point will be reflected back from the mirror or refracted by the lens parallel to the optical axis.
- A ray arriving to the centre of the mirror will be reflected back symmetrically to the optical axis. In case of a lens its direction will not be altered.

In the following examples we will construct the images formed by lenses, but the projections of mirrors may be constructed in an equally simple manner.

- Figure 8.6 demonstrates the case where an object is placed in front of a convex lens, and the object distance is larger than the focal distance. The refracted rays of light are convergent: they cross each other on the other side of the lens. If we were to place a screen to this position a clear image of the object would appear on it. Therefore this kind of image is known as a real image.

It must also be noted that the AMB and CMD triangles are similar triangles: the lengths of their corresponding sides are proportional to each other. Based on this we may determine the magnification of the image from the ratio of the image distance and the object distance:

$$M = \frac{AB}{DC} = -\frac{AB}{CD} = -\frac{AM}{CM} = -\frac{i}{o} \quad (8.42)$$

where o is the object distance, and i is the image distance. The negative sign signifies that the image stands upside down.

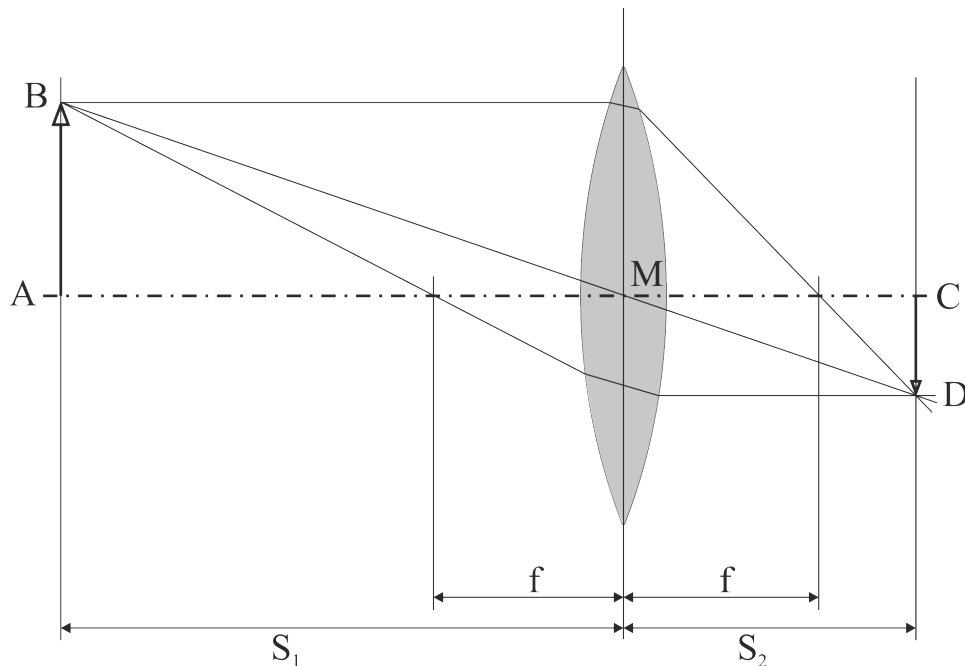


Figure 8.6: Construction of the image formed by a converging lens if the object is farther away from the lens than its focal distance

- If the object is placed precisely into the focal point of a convex lens all rays of light originating from it will be refracted parallel to the optical axis, and no image will be formed.
- In figure 8.7 the object is placed closer to a convex lens than its focal point. In this case the rays originating from the object are divergent even after they were refracted by the lens. This means they will never cross each other, and no real image will be formed. (There is no position in the optical system, where a clear image would appear on a screen.)

It must be noted however that if we extend the refracted rays backwards, these extensions (marked by dotted lines on the figure) will cross each other behind the object. In other words for an observer looking into the optical system from the right it will seem like the refracted rays are originating from that point: the observer will see a so called virtual image behind the lens.

The magnification can be calculated again by equation (8.42). Since in this case the image distance is negative, the two negative signs cancel out each other, and the magnification becomes positive. (The image stands upright.) As in this case the image always appears behind the object, the image distance is always larger than the object distance, and the image is always magnified.

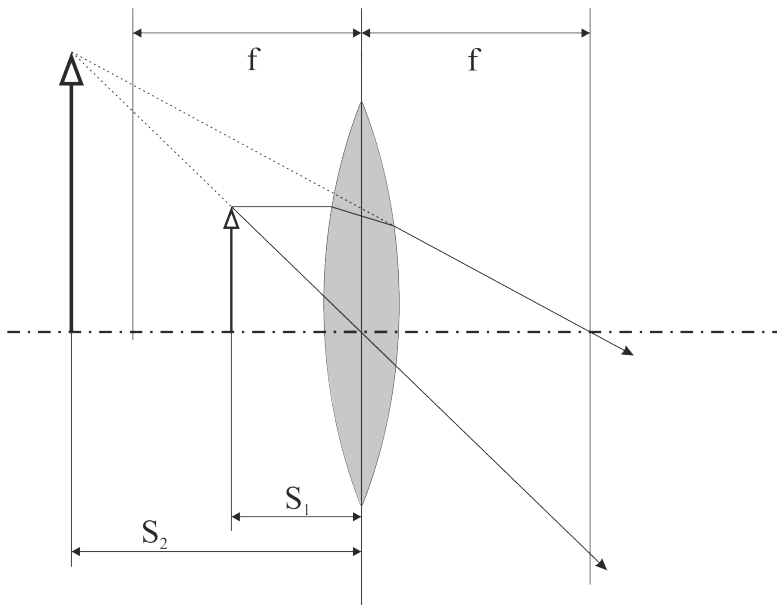


Figure 8.7: Construction of the image formed by a converging lens if the object is closer than its focal distance

- If we place an object in front of a concave lens (also known as a diverging lens or negative lens), the refracted rays will be even more divergent than the original rays. Therefore the image is virtual. A diverging lens can never form a real image regardless of the position of the object. As the image distance is always negative for a virtual image, the magnification is always a positive number, therefore the image stands upright.

8.5 Aberrations

In the deductions of the previous sections we have made several approximations. This means that although the resulting formulas are close to being true in the paraxial approximation, they are never exact. Also the shapes of mirrors and lenses can never be completely precise: there are always minor deviations from the ideal due to the inaccuracies of the manufacturing process. (The optical components may change their parameters even after manufacturing due to mechanical stress or temperature differences.) The materials used to build the optical systems may scatter light or behave differently at different wavelengths. All these deviations from the ideal behaviour are referred to as aberrations.

- Chromatic aberration

Although during the discussion of lenses we have assumed that their index of refraction is a constant, in practice this is not true. The index of refraction of practical

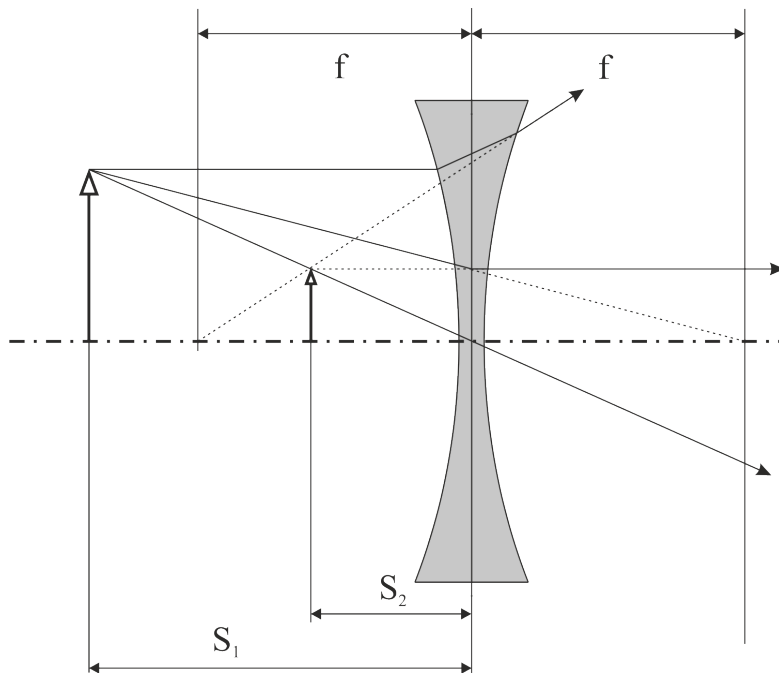


Figure 8.8: Construction of the image formed by a diverging lens

materials depends on the wavelength of light. Since the focal length changes with the index of refraction, lenses have slightly different focal distances for different wavelengths. The result is that we cannot get a clear image for all colours at the same time. For example if we set up the optical system to give a clear image in blue light, images in all other colours will be slightly blurred, and we will see coloured outlines around objects on the image.

There are two possible solutions to this problem. Reflectors have no chromatic aberration, therefore by replacing lenses by mirrors, this aberration can be eliminated. However this is not always practical. In certain optical systems the use of lenses is advantageous. Therefore opticians have developed so called achromatic lenses or acromats, in which lenses of different materials are assembled together to form a compound lens. On their own, each of the components have chromatic aberration, but since their materials are different, their chromatic aberrations are also different. If designed properly these different aberrations of the components may cancel out each other, and the aberration of the compound lens can be minimised.

- Spherical aberration

Although in the previous sections we have considered spherical reflectors and refractors, it was mainly for the sake of simplicity. Our approximations are valid only for rays of light very close to the optical axis. It can be shown, that the focal

distance of a spherical lens or mirror changes with the distance from its centre. (Marked by h on the figures.) This means, that a ray of light arriving to the edge of a spherical mirror will be focused to a different point than another which arrives close to the centre. In other words we can never get a clear image with a spherical lens or mirror.

It can be proven that the exact shape of the mirror should be a paraboloid instead of a sphere segment to avoid this kind of aberration. Unlike a spherical mirror a parabolic mirror can focus large diameter beams into a single point. (As long as the beam arrives parallel to the optical axis.)

- Coma

Comatic aberration is an inherent property of parabolic mirrors. Although they can focus a wide beam of light into a single point, this is possible only if the beam arrives parallel to the optical axis. Rays arriving from off-axis directions will not be focused to the single point. The result is that images of objects that are not in the centre of the field of view are going to be blurred. The image of an off-axis point source (such as a star which is not in the centre of the field of view) is not a single point, but a wedge-shaped smear, resembling the coma of a comet, hence the name.

The effect may be reduced by introducing appropriately shaped correction plates or correction lenses into the optical system (as in the case of Maksutov- or Schmidt telescopes), or by replacing some of the parabolic mirrors in the system by hyperbolic mirrors.

- Field curvature

Photo-plates or digital image sensors, such as CCD or CMOS matrices are usually flat. But most optical systems project the image of objects onto a slightly curved surface. (Imagine that an object is placed in front of an “ideal” lens. If the distance of the centre of the flat sensor from the centre of the lens is exactly the image distance, the system will give a clear image in the vicinity of the optical axis. But off-axis parts of the flat sensor are at a larger distance from the centre of the lens, therefore they are going to be out-of-focus.) In other words either the edges or the centre of the image is going to be slightly blurred if we use a flat image sensor. The phenomenon is called Petzval field curvature. It can be remedied by building appropriately curved image sensors. This was relatively easy to achieve with traditional photo-films, as they were flexible, and could be stretched to the appropriate shape. Modern semiconductor based sensors are too rigid and fragile for such techniques, but in some instances (such as in the case of the Kepler space telescope where a large field of view was required) large mosaic-like image sensor arrays may be constructed to compensate for field curvature. Also the lenses in

modern cameras are designed to have larger focal distances for off-axis rays to minimise field curvature.

- Astigmatism

Practical optical systems never have perfect rotational symmetry. The curvature of lenses and mirrors is usually slightly stronger in one direction, and they their axis can never be perfectly aligned with the axis of the system either. This means then lenses and mirrors have slightly different focal distances in two perpendicular directions (such as in the horizontal and vertical direction). If we place the screen or the image sensor into one of these focal planes, the image is going to be blurred in the other direction. There is no position where we could get an image which is perfectly clear in both directions. (Placing the sensor between the two focal planes smears the image in both directions. . .) Such aberrations created by errors in the shape of lenses and mirrors, or by misalignments are usually referred to as astigmatism.

Chapter 9

Wave optics - Gábor Dobos

9.1 Young's double slit experiment

In the previous chapter we have discussed optical systems using geometrical optics, and ignored any wavelike behaviour. But as we have seen earlier, light is a form of electromagnetic waves, which means that in certain cases geometrical optics won't be sufficient to describe its behaviour. The first such experiment was presented by Thomas Young in 1802.

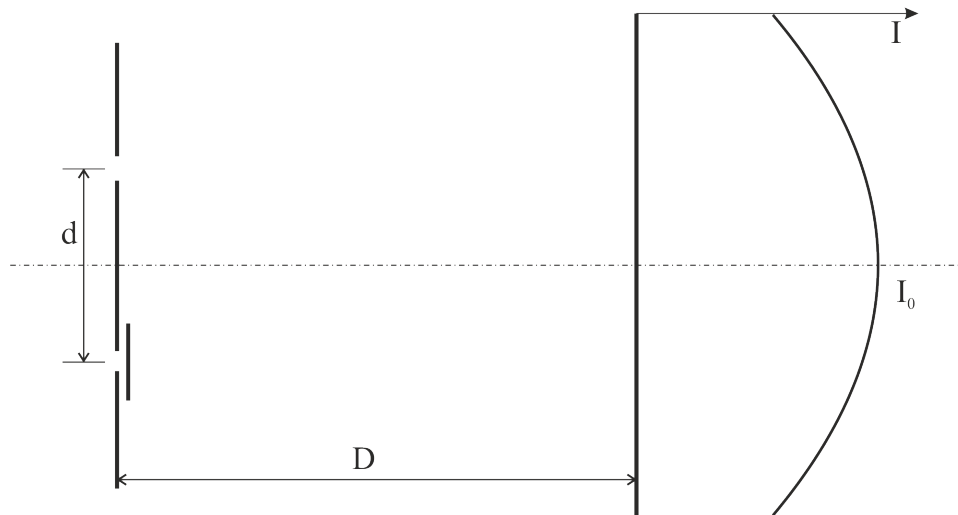


Figure 9.1: Illumination of the screen from a single slit (Image is not to scale)

Young put a thin plate with two small holes on it in front of a light source, and observed the light passing through it on a screen placed a few metres behind the plate. When either one of the holes was covered, and light was allowed to pass through only one of the holes, the whole screen was illuminated (Figure 9.1). Based on geometrical optics,

one would expect, to get the same pattern with doubled intensity when both holes are uncovered. (The holes are so close to each other, that light passing through either one of them gives virtually the same distribution on the screen.) But the experiment showed that when light was allowed to pass through both holes simultaneously a pattern of bright and dark rings have appeared on the screen, and the intensity of the bright spot in the centre has not doubled but quadrupled. These results cannot be explained by geometrical optics. To give a proper description of the phenomenon we have to consider light as a wave once again.

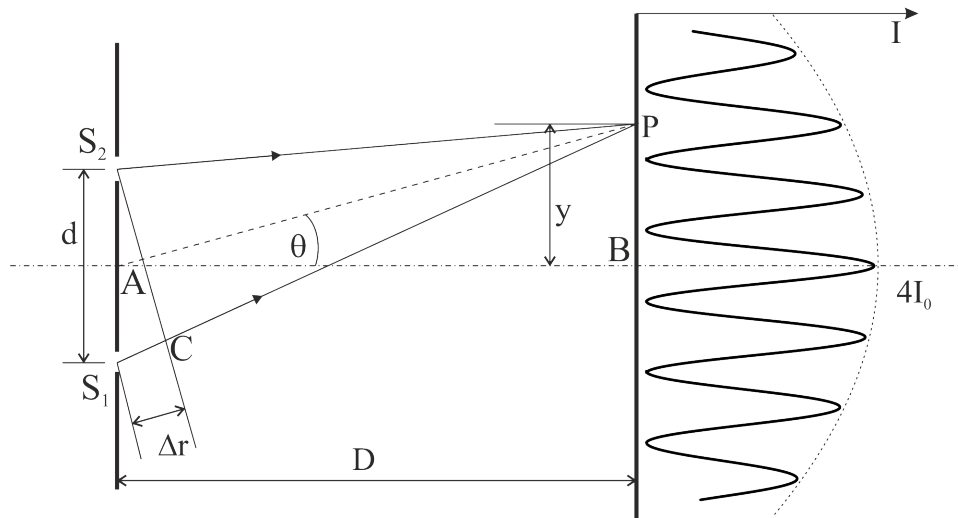


Figure 9.2: When light is allowed to pass through both slits simultaneously (double slit experiment) a diffraction pattern consisting of a series of bright and dark lines appears on the screen (Image is not to scale)

Later Young has repeated his experiment with thin slits (hence the name: double slit experiment) instead of small circular holes, to increase intensity. In this case a series of bright and dark lines appear on the screen instead of rings, but the principle is the same. Electromagnetic waves leaving the source reach both slits, and pass through them. For an observer on the other side of the plate it seems like both slits are emitting electromagnetic waves of the same frequency and wavelength. To calculate the light intensity at a given point of the screen we have to consider the superposition of these two waves.

The situation is very similar to the superposition of two harmonic oscillations of the same frequency. As we have seen in the previous semester the superposition of two such oscillations is another oscillation of the same frequency, whose complex amplitude is the sum of the complex amplitudes of the individual oscillations. This means, that the amplitude of the oscillation depends not only on the amplitudes of the individual oscillations, but also on their phase difference. The amplitude is maximal when the oscillations are in phase (their phase difference is 0 or integer times 2π), and minimal if

they are in opposing phases (the phase difference is π , or 3π , or 5π , etc...).

In case of the double slit experiment we get a bright area on the screen where the two waves reach it in the same phase (constructive interference), and a dark area where they meet in opposing phases (destructive interference). On figure 9.2 light coming from S_1 has to travel a longer distance to reach point P, therefore it will be late compared to light coming from S_2 . To calculate the phase difference we have to determine this optical path difference. Since the S_1S_2C triangle is similar to the PAB triangle the S_1S_2C angle is equal to θ . Therefore:

$$\Delta r = d \sin \theta \quad (9.1)$$

Each wavelength extra optical path increases the phase of a wave by 2π , therefore the phase difference is:

$$\Delta \phi = 2\pi \frac{\Delta r}{\lambda} = 2\pi \frac{d}{\lambda} \sin \theta \quad (9.2)$$

It must be noted, that the distance of the screen from the slits (D) is several orders of magnitude larger than the distance of the bright fringes from the centre of the screen (y), therefore the θ angle is very small. For small angles $\sin \theta \approx \tan \theta = \frac{y}{D}$, thus:

$$\Delta \phi = 2\pi \frac{d}{\lambda} \sin \theta \approx 2\pi \frac{d}{\lambda} \frac{y}{D} \quad (9.3)$$

The screen will be bright where the two waves meet in the same phase, thus their phase difference is $m2\pi$, where m is an integer number. (Or in other words their optical path difference is an integer times the wavelength.) Therefore the criterion of constructive interference is:

$$2m\pi = 2\pi \frac{d}{\lambda} \sin \theta \approx 2\pi \frac{d}{\lambda} \frac{y}{D} \quad (9.4)$$

$$m\lambda = d \sin \theta \approx d \frac{y}{D} \quad (9.5)$$

Based on equation (9.5) we may determine the position of the bright fringes on the screen:

$$y_m \approx \frac{mD\lambda}{d} \quad (9.6)$$

Therefore the distance between the m^{th} and $(m+1)^{th}$ bright fringe is:

$$\Delta y = y_{m+1} - y_m \approx \frac{(m+1)D\lambda}{d} - \frac{mD\lambda}{d} \quad (9.7)$$

$$\Delta y \approx \frac{D\lambda}{d} \quad (9.8)$$

Thus the bright areas on the screen are equidistant.

In a similar manner we may also determine the positions, where the screen will be dark. For this the two waves have to meet in opposing phases (their phase difference must be $(2m + 1)\pi$, where m is an integer number) to cancel out each other. Therefore the criterion of destructive interference is:

$$(2m + 1)\pi = 2\pi \frac{d}{\lambda} \sin\theta \approx 2\pi \frac{d}{\lambda} \frac{y}{D} \quad (9.9)$$

$$\left(m + \frac{1}{2}\right) \lambda = d \sin\theta \approx d \frac{y}{D} \quad (9.10)$$

The phenomenon is usually referred to as diffraction. The bright areas are called diffraction lines or fringes and the integer number m is the order of the diffraction.

Based on the above description we may also understand why the intensity of the bright areas has quadrupled instead of just doubling. In case of constructive interference the two waves meet in the same phase, thus the amplitude doubles. But we have also deduced earlier that the intensity of light is proportional to the square of the amplitude. Thus, doubling the amplitude quadruples the intensity.

This may seem to violate the principle of energy conservation (How can the intensity quadruple? Where is the extra light coming from?), but it does not. Don't forget, that other areas of the screen that were illuminated when light was allowed to pass through only one hole, became dark due to destructive interference. It is important to understand that neither constructive nor destructive interference can create or destroy energy. When we uncover the second slit the total amount of light reaching the screen doubles, but its distribution also changes. Diffraction "redirects" some of the light from the dark areas to the bright ones without changing the total power that reaches the screen. Therefore diffraction (and interference phenomena in general) does not violate the principle of energy conservation, it only changes the distribution of light.

9.2 Coherence

Equation (9.6) shows that the position of diffraction lines depends not only on the distance of the slits from each other, and from the screen, but also on the wavelength of the light. This means that if we repeat the experiment using a light source of different wavelength, the position of the diffraction lines also changes. From equation (9.6) we may determine the shift in the position of diffraction lines due to a shift in wavelength:

$$\Delta y_m \approx \frac{mD\Delta\lambda}{d} \quad (9.11)$$

The problem is that practical light sources are never completely monochromatic: their spectrum always includes different wavelengths. This means that diffraction patterns

belonging to different wavelengths in the spectrum of the light source appear on the screen superimposed on each other. Because of the m multiplier on the right-hand side of equation (9.11) higher order diffraction lines suffer larger shifts in position due to the same shift in wavelength. This means that while the first order diffraction lines may be clearly resolved the 15th order bright line of one wavelength may very well coincide with the 16th order dark line of another wavelength. In other words higher order diffraction lines get smeared more easily.

The result is that only a finite number of lines are visible on the diffraction pattern: the lower order lines are usually well resolved, but as the order of diffraction increases the pattern gets more and more smeared and after a point the lines cannot be distinguished any more. Of course the number of visible lines depends on the light source: the more monochromatic it is, the smaller the shifts in line positions will be, and the more lines will appear clearly resolved.

The maximum optical path difference at which interference can still be observed (the diffraction lines are still resolved) is called the coherence length of the light source

$$L = \frac{c}{n\Delta f} = \frac{\lambda^2}{n\Delta\lambda} \quad (9.12)$$

where c is the speed of light in vacuum, n is the index of refraction of the medium in which the experiment is done (thus $\frac{c}{n}$ is the speed of light in the medium), f is the frequency of the light, and λ is its wavelength. Δf and $\Delta\lambda$ are the spectral width of the light source in frequency and in wavelength, respectively.

But coherence means slightly more than just a monochromatic spectrum. Diffraction (as any other interference phenomena) depends on the phase of the interfering waves. A more precise definition of coherence length states that it is the propagation distance over which a wave maintains its coherence, or in other words its phase remains predictable. One may appreciate that this definition is very similar to the previous one. Waves of different wavelengths complete a different number of oscillations while travelling the same distance, thus on arrival their phases will be different. The longer they have to travel and the larger the wavelength difference is, the larger phase difference they accumulate. The phase of waves emitted from a non-monochromatic source becomes less and less predictable the longer they need to travel, and the less monochromatic the source is. Coherence length is the distance over which phase remains predictable, or in other words diffraction patterns of different wavelengths does not smear each other.

Another way to characterise the light source is to give its coherence time. By definition coherence time is the time over which a wave might be considered coherent, or in other words it maintains a predictable phase. The coherence time can be calculated by dividing the coherence length by the velocity of propagation:

$$\tau = \frac{1}{\Delta f} = \frac{\lambda^2}{c\Delta\lambda} \quad (9.13)$$

It must be noted that the invention of lasers was a very important achievement in the history of science and technology, as they are the most coherent (and most monochromatic) man-made light sources. While the coherence lengths of traditional light sources extend only to a couple of wavelengths, the coherence length of even a simple multi-mode laser is in the range of dozens of centimetres, while that of single mode lasers may be hundreds of meters, and the coherence length of fibre lasers may exceed a hundred kilometres!

9.3 Multiple slit diffraction

A similar experiment may be carried out using not two, but multiple slits. For the sake of simplicity let us consider a triple slit experiment! According to figure 9.3 the optical path difference, and the phase shift between waves origination from neighbouring slits may be calculated in a similar fashion as before:

$$\Delta\phi = 2\pi \frac{d}{\lambda} \sin\theta \approx 2\pi \frac{d}{\lambda} \frac{y}{D} \quad (9.14)$$

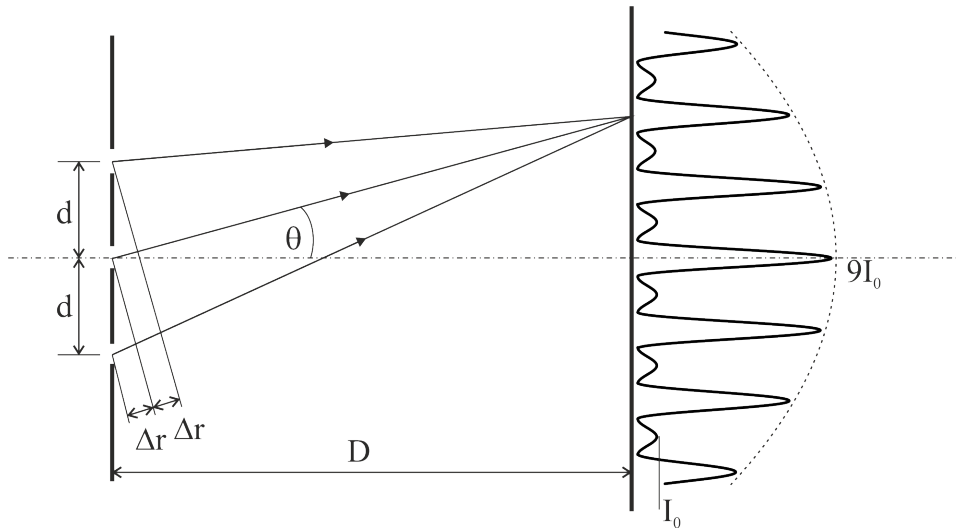


Figure 9.3: In case of a triple slit experiment the position of the major peaks are the same as in the double slit experiment, but they are narrower and more intense. Also a minor peak appears between each pair of major ones.

The main difference is that in this case we have to consider the superposition of not two but three waves. Naturally, the resulting complex amplitude is going to be the sum of the complex amplitudes of all three waves. In the centre of the screen, all three

waves meet in phase, thus all three phasors point in the same direction (figure 9.4) and consequently the amplitude triples (compared to the amplitude of the wave coming from a single slit), while the intensity of the light increases ninefolds.

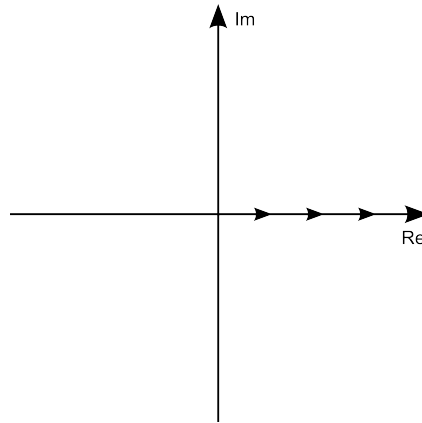


Figure 9.4: The waves originating from the three slits reach the centre of the screen in the same phase, thus the three phasors representing them point in the same direction. This triples the amplitude and increases the intensity ninefolds.

As we move away from the centre of the screen, the optical path difference (and the phase difference) between the waves increases, and the intensity decreases. (The three waves are no longer in phase, thus the phasors do not point in the same direction.) The intensity becomes minimal, when the phase difference reaches $2\pi/3$ or 120° . (Figure 9.5) In this case the three phasors cancel out each other, thus the amplitude and the intensity of light is zero.

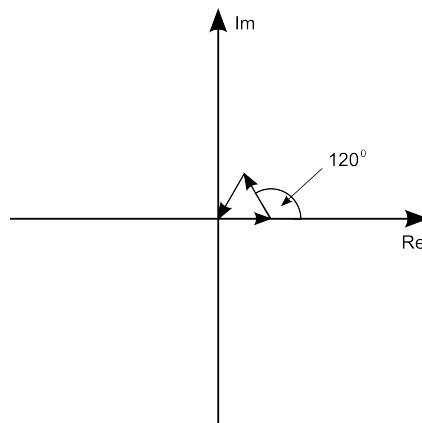


Figure 9.5: When the phase difference between the waves is 120° they cancel out each other.

If we move further away from the centre the phase difference keeps increasing, and the three waves do not cancel out each other anymore, thus the intensity starts increasing again. We may detect a minor maximum, when the phase difference reaches π . (Figure 9.6) In this case two of the phasors point in one direction and third in the opposite direction. Thus the amplitude is the same as the amplitude of a single wave, and the intensity is equal to the intensity we may detect when the screen is illuminated only by a single slit.

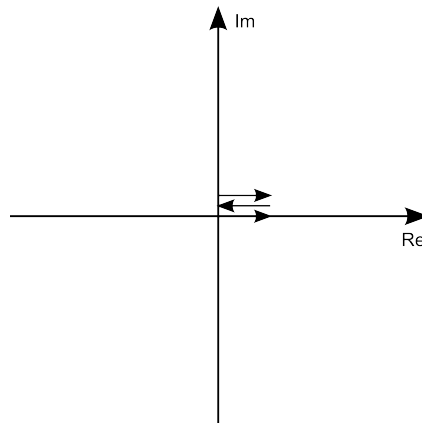


Figure 9.6: When the phase difference is 180° two phasors point in one direction, and the third in the opposite direction. Thus the amplitude and the intensity is the same as if the screen was illuminated through a single slit.

Further away from the centre the intensity keeps decreasing again until the phase difference reaches $4\pi/3$ or 240° . (Figure 9.7) In this case the three phasors cancel out each other yet again, and the intensity is zero. After this, the amplitude increases with the phase difference, until at 2π phase difference the phasors point in the same direction yet again (figure 9.4), thus the amplitude is three times the amplitude of a single wave, and the intensity is nine times higher than what we may detect when the screen is illuminated through a single slit.

If we keep moving away from the centre this pattern repeats itself over and over again. Major maxima appear where the phase difference between waves originating from neighbouring slits is an integer times 2π . (Or in other words the optical path difference is an integer multiple of the wavelength.) Thus the position of these major maxima is the same as in case of the double slit experiment. Since in these cases the amplitude is thrice the amplitude of a single wave, these major maxima are nine times more intensive than the illumination from a single slit.

Halfway between each major maxima pair is a minor maximum (where two phasors point in the same direction, and one in the opposite direction). The maxima are separated by dark regions, where the phase difference is either 120° or 240° , and the three waves cancel out each other.

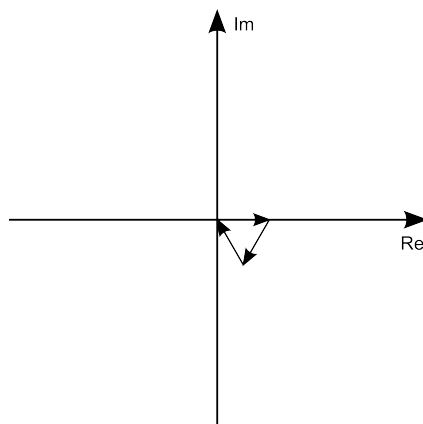


Figure 9.7: When the phase difference reaches 240° , the three waves cancel out each other yet again.

Following this scheme, it is easy to see, what happens when we increase the number of slits. Major maxima appear always at the same positions given by equation (9.6), where the phase difference between waves originating from neighbouring slits is an integer times 2π . In these cases all phasors point in the same directions, thus the amplitude is proportional to the number of slits, and the intensity is proportional to its square.

Between the major maxima, a number of minor maxima appear. While in case of a double slit experiment we get no minor maxima, in case of a triple slit experiment we get a single minor maximum between major ones. In case of a four slit experiment we get two minor lines, in a five slit experiment, we get three, and so on. The number of minor lines between the major ones is always the number of slits minus two.

It is also worth noting, that with the increase of the number of slits, not only the intensity of the major lines increases, but they are also getting narrower. Using a very large number of slits results in very narrow diffraction lines (compared to the distance of the lines). Thus the lines are well separated. This, combined with the fact that line positions depend on the wavelength of light, makes a plate with a very large number of slits on it an ideal dispersive element. In other words it can be used to separate light of different wavelengths.

The first optical spectrometers used prisms to separate different wavelengths.¹ But most modern optical spectrometers use so called diffraction gratings for this purpose. These act like a plate with a large number of slits on it in the multi slit experiment. They diffract different wavelengths in different directions. As these gratings have a very

¹The refractive index of glass depends on the wavelength of light. Thus different colours are refracted into slightly different directions when light enters or leaves the prism through a face that is not perpendicular to its direction. Therefore a prism can be used to resolve white light to the colours of the rainbow as it was demonstrated by Newton in the late 1660's. (In fact, actual rainbows are created in a similar manner when sunlight is refracted by small raindrops floating in the air after a rain...)

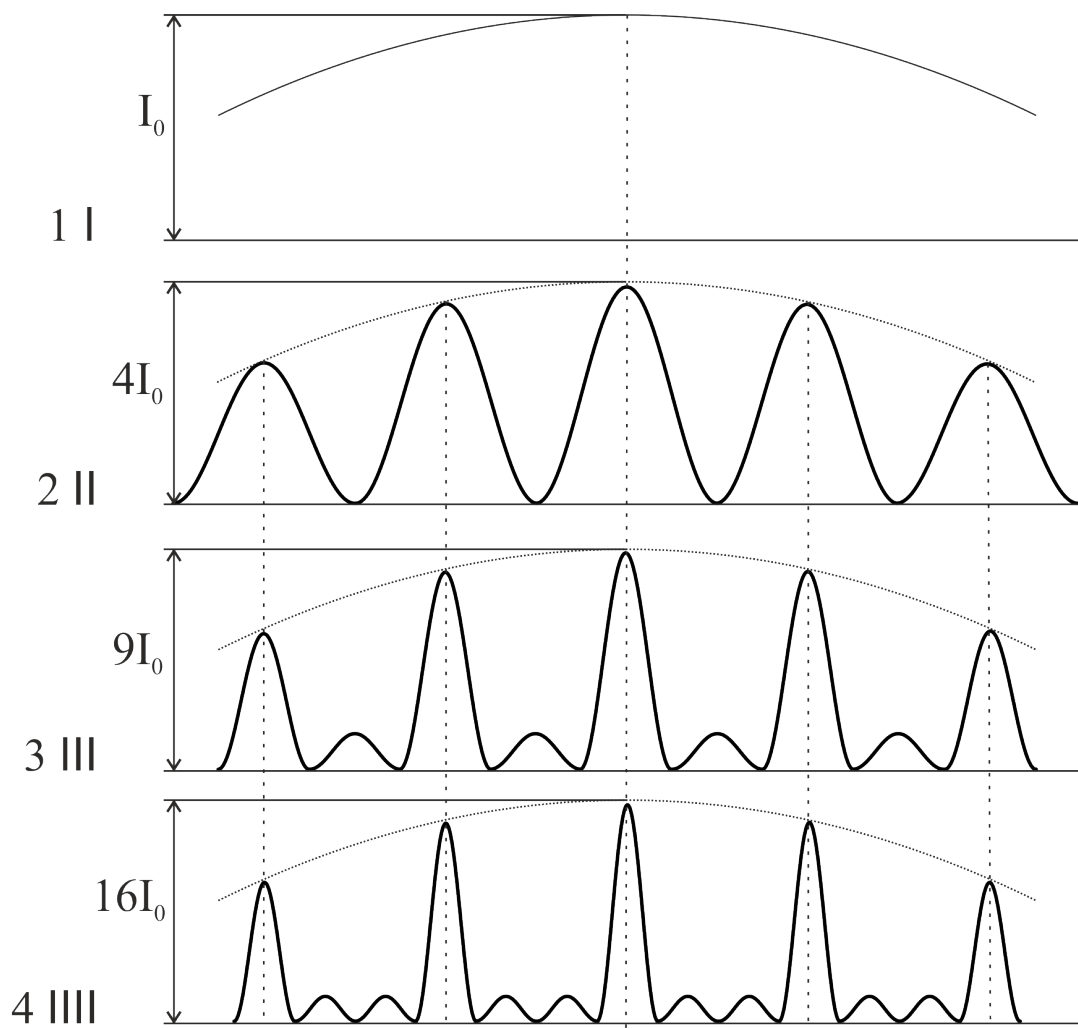


Figure 9.8: The intensity of the major lines increases by the square of the number of slits. They also get narrower and a series of minor peaks appear between them.

large number of “slits”² each diffraction line is very narrow, thus the spectrum of light falling on the grating is well resolved. It must be noted however, that these diffractive spectrometers may generate certain artefacts in the spectrum. (For example the first order diffraction line of a long wavelength may overlap with the second order diffraction line of a shorter wavelength...)

9.4 Fraunhofer diffraction

Experiments show that under certain circumstances diffraction patterns may be obtained even if we use only a single slit. This kind of diffraction phenomena is referred to as Fresnel- or Fraunhofer diffraction depending on the divergence of the incident light. In case of Fresnel diffraction the light source is relatively close to the slit, the incoming rays are divergent, and light reaches the slit in form of spherical waves. On the other hand if the light source is relatively far from the slit, the divergence of the incoming rays becomes negligible, and light reaches the slit as a plane wave. Such cases are referred to as Fraunhofer diffraction.

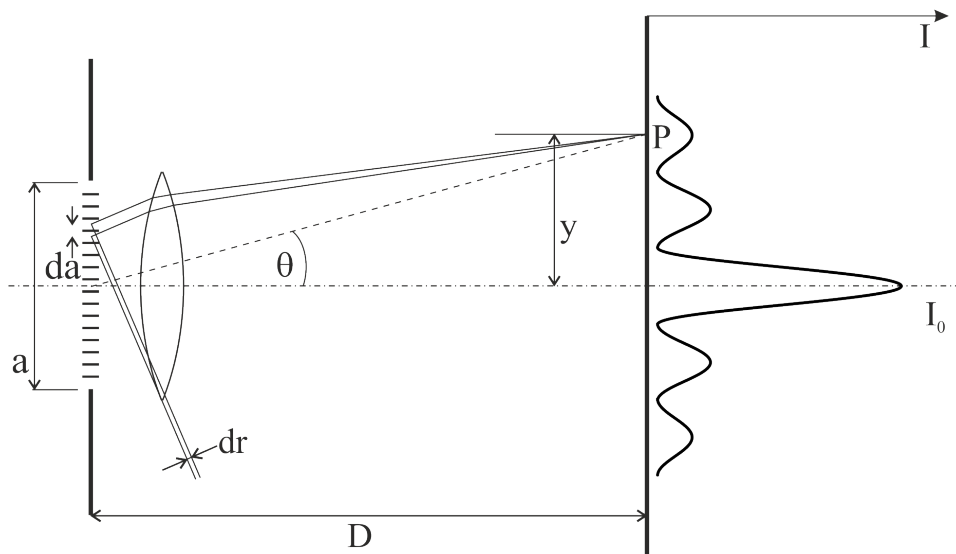


Figure 9.9: Diffraction on a single wide slit.

The two phenomena are similar to each other, and for the sake of simplicity we will discuss only Fraunhofer diffraction. In this case not only the incoming waves are plane

²It must be noted, that modern optical gratings are often so called reflective gratings, that are made by creating a very dense line pattern on a reflective surface. Due to the pattern, light can be reflected from some areas (that act like the slits in the traditional multislit experiment), and not from others. An observer looking at the reflection of a light source will perceive it as if looking at it through a plate containing a large number of narrow slits.

waves, but the diffracted waves, too. This means, that the screen should either be very far from the slit, or we should place a lens in front of the slit. (A lens focuses rays heading into the same direction into a single point on the screen which is placed into its focal plane. Thus the lens projects each diffraction direction to a different point of the screen.)

To understand the phenomena imagine that we divide the single slit to a large number of small segments. Each of these segments acts as a light source, “emitting” electromagnetic waves with a constant phase displacement from each other. The diffraction pattern is produced by the interference of these waves, and just like in case of the multi slit experiment the amplitude at a given point of the screen may be calculated as sum of the phasors representing the complex amplitudes of the waves. From figure 9.9 the phase difference between waves from neighbouring segments is:

$$d\phi = 2\pi \frac{dr}{\lambda} = 2\pi \frac{da \sin \theta}{\lambda} \approx 2\pi \frac{da}{\lambda} \frac{y}{D} \quad (9.15)$$

Where a is the width of the slit, and da is the width of each segment. Since the phase difference is constant, and the amplitudes are the same (the slit is illuminated homogenously, and the sizes of the segments are identical), the phasors representing the complex amplitudes of the waves form a circular arc. The angle of this arc is:

$$\phi = \int d\phi = 2\pi \frac{a \sin \theta}{\lambda} \approx 2\pi \frac{a}{\lambda} \frac{y}{D} \quad (9.16)$$

The total amplitude in the θ direction may be calculated from the OPQ triangle on figure 9.10:

$$A_\theta = 2R \sin \left(\frac{\phi}{2} \right) \quad (9.17)$$

The only unknown in equation (9.17) is the parameter R . This may be determined from the length of the circular arc, which is formed by the phasors representing the complex amplitudes of the waves. As the magnitude of these amplitudes is constant, the total length of the arc is also constant. (It does not depend on ϕ .) Let us mark the intensity in the centre of the screen by I_0 , and the corresponding amplitude by A_0 . Since all waves reach this point in the same phase, all phasors point in the same direction. Thus the length of the “arc” is equal to A_0 . From this:

$$A_0 = R\phi \Rightarrow R = \frac{A_0}{\phi} \quad (9.18)$$

Substiuting this into equation (9.17) gives:

$$A_\theta = 2 \frac{A_0}{\phi} \sin \left(\frac{\phi}{2} \right) = A_0 \frac{\sin \frac{\phi}{2}}{\frac{\phi}{2}} \quad (9.19)$$

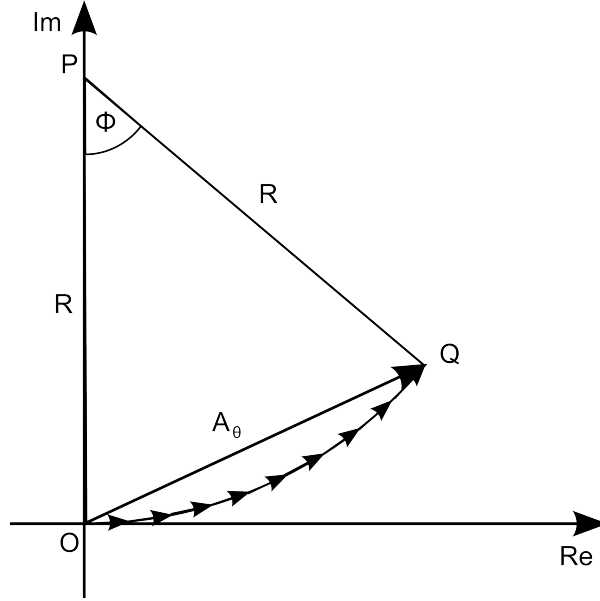


Figure 9.10: The phasors representing the complex amplitudes of waves originating from different segments of the slit from a circular arc.

The intensity is proportional to the square of the amplitude, thus

$$I_{\theta} = I_0 \left(\frac{\sin \frac{\phi}{2}}{\frac{\phi}{2}} \right)^2 \quad (9.20)$$

The screen is dark, where $\sin \frac{\phi}{2} = 0$, thus the criterion of destructive interference is:

$$\frac{\phi}{2} = m\pi \quad (9.21)$$

$$\phi = 2m\pi \quad (9.22)$$

$$2\pi \frac{a \sin \theta}{\lambda} = 2m\pi \quad (9.23)$$

$$a \sin \theta = m\lambda \quad (9.24)$$

where m is an integer number.

Note, that although equation (9.24) takes the same form, as equation (9.5), here a marks the width of the single slit, and not the distance of the slits. Also, in case of Fraunhofer diffraction, the equation gives the position of the dark areas, whereas in case of the multislit experiment it gave the position of the bright areas.

The significance of the phenomena is that it determines the maximal resolution achievable by an optical system. Every such system (let it be a telescope, a microscope or any other optical instrument) has an aperture where light enters the system. This aperture acts like the slit in the deduction above. Light passing through it will undergo diffraction, changing its direction of propagation and smearing the image.

Imagine that we have an optical system which projects the image of an object to a screen. Even if light is coming from a point source it can never be projected into a single point, instead it will produce a diffraction pattern like the one on figure 9.9. Even the central maximum has a finite width, thus if two points of the object are too close to each other, the corresponding diffraction patterns will overlap on the screen, and we won't be able to distinguish them. In other words even if all aberrations of the optical system are negligible, and everything is in perfect focus, the achievable resolution is limited by diffraction. Therefore such optical systems are referred to as diffraction limited optical systems.

From the deduction above, it is obvious that the achievable resolution depends both on the wavelength of the light, and the size of the aperture (or entrance slit), through which light enters the optical system. The shorter the wavelength, and the larger the aperture, the better the resolution is going to be. According to the Rayleigh criterion two points on the image are well resolved, if the central maximum of the diffraction pattern of one of them, is not closer to the other, than the first minimum of its diffraction pattern.

It must be noted, that in case of circular apertures equation (9.24) takes a slightly different form:

$$D \sin \theta = p_m \lambda \quad (9.25)$$

where D is the diameter of the aperture and $p_1 = 1.22$, $p_2 = 2.233$, $p_3 = 3.238$, etc... Usually the diffraction angles in practical optical systems are so small, that $\sin \theta$ may be approximated by θ . Using these, the Rayleigh criterion may be stated in a mathematical form:

$$\theta = \frac{1.22 \lambda}{D} \quad (9.26)$$

Thus the angular resolution of a diffraction limited optical system is proportional to the wavelength of the light and inversely proportional to the diameter of the aperture. Based on equation (9.26), it is easy to understand why astronomers prefer large telescopes. It's not only because a larger main mirror may collect more light and produce a brighter image: the maximal resolution of the optical system also depends on the size. It must be noted, that while the brightness of the image depends on the area of the main mirror, the resolution depends on the distance between its most distant edges. This means, that if parts of the main mirror are missing, it decreases only the brightness of the image, and has little or no effect on the resolution (as long as the maximal distance of the remaining parts are the same, as in case of a full mirror).

This means, that in order to achieve a high resolution, we don't have to build a complete large mirror. It is enough to build its most distant sections, since the missing parts won't affect the resolution, only the brightness of the image. In practice this is realised by building several smaller telescopes, that can move together, and point to the same section of the sky. Light collected by these telescopes is united, thus making it possible for the telescopes to function as one large telescope. Such systems are usually referred to as astronomical interferometers, and the distance of the parts is called the baseline of the interferometer.

The first such system was the Michelson stellar interferometer. It had such a high resolution that in 1920 it made it possible to measure the diameter of distant stars for the first time. Today many modern telescopes are capable to work in interferometric mode including the Keck Observatory in Hawaii, the Large Binocular Telescope in Arizona, and the Very Large Telescope, operated by the European Southern Observatory in Chile. All of these systems have several telescopes that may be linked together to function as an astronomical interferometer, greatly improving their resolution.

It must also be noted that since the 1940's radio telescopes are also routinely used in interferometric mode. That is the reason why large radiotelescope systems usually have several smaller dishes instead of a single large one. The individual radio telescopes do not even have to be at the same location. Measurement data recorded by radio telescopes thousands of kilometres away from each other may be transmitted to a single centre for processing, forming in effect a single large radio telescope. The largest such system in operation today is the Very Long Baseline Array whose telescopes are located all over the United States, forming an astronomical interferometer with a maximal baseline of 8611 kilometres.

But the Rayleigh criterion affects not only astronomy. Since the wavelength of visible light is several hundred nanometres, traditional optical microscopes usually have a resolution of only a few microns, or a few hundred nanometres at best.³ This is one of the reasons why scientists prefer electron microscopes. Electron beams accelerated to a suitably high energy have wavelengths several orders of magnitude smaller than visible light, making it possible for certain types of electron microscopes to achieve atomic resolution, where individual atoms of the sample are visible on the image.

The Rayleigh criterion also poses a serious problem to the semiconductor industry. Since integrated circuits are manufactured using photolithography, the achievable resolution is also limited by diffraction phenomena. Currently (2013) the most advanced

³It must be noted that in microscopy the Rayleigh criterion is given in a slightly different form: $R = \frac{1.22\lambda}{NA}$, where R is the smallest distance resolved by the microscope, $NA = n_m \sin\alpha$ is called the numerical aperture of the lens, α is the half-angle of the maximum cone of light that can enter the lens and n_m is the refractive index of the medium surrounding the lens. Since the numerical aperture depends on the refractive index of the medium, spatial resolution may be improved by immersing the optical system in a refractive liquid medium (for example: water). Hence such systems are referred to as immersion optics, and they are widely used in modern semiconductor manufacturing technology.

semiconductor devices use 22 nm technology, which means that the smallest feature of the device is approximately 22 nm. These devices are manufactured using 193 nm UV light. The problem is that decreasing the wavelength of light to improve resolution has become increasingly harder in the past few years. Currently it is expected, that the next technological step (14 nm devices - coming into production in 2015) will be achievable using a more advanced version of the current technology, but the step after that (10 nm, est. 2017) will require a significantly different manufacturing process.

Due to the difficulties posed by extreme ultraviolet lithography, currently it is expected that 10 nm technology will use the same 193 nm wavelength as today's technology, and the improvement in resolution will be made possible by other means, such as multiple patterning. In this case high resolution patterns are achieved by the overlapping of several lower resolution patterns. (The overlapping area might be considerably smaller than the resolution of each pattern, thus increasing the overall resolution...)

9.5 Thin layer interference

Thin layer interference is a type of interference that we have all encountered in our daily life. This is what makes soap bubbles, and oil spills after a rain sparkle in rainbow colours. The reason of the phenomena is that sunlight contains all colours of the rainbow, and reflection from such thin layers depends on the ratio of the layer thickness and the wavelength of the light. A layer of a given thickness may reflect one wavelength but not another, giving it a distinctive hue. Consequently, as the thickness of the layer changes from point to point, so does its hue. Since both oil spills and soap bubbles have varying thickness, they will sparkle in rainbow colours.

To better understand the phenomena let us consider the reflection of light from the wall of a soap bubble. (This wall is basically a very dilute aqueous solution, thus its index of refraction is similar to that of water.) Some of the light arriving to the air/water interface will be reflected back, while the rest will be transmitted into the water. The transmitted light will reach the water/air interface on the other side, and again, some will be reflected back, while the rest will be transmitted, and leave the wall of the bubble. The waves reflected back from the air/water and from the water/air interface may both reach an observer, and interfere with each other. Depending on the phase difference, the interference may either enhance the intensity (constructive interference), or the waves may cancel out each other (destructive interference).

To determine the phase difference, first we have to determine the optical path difference between the rays reflected back from the two interfaces. On figure 9.11 the ray that enters the wall, and bounces back from the water/air interface travels from point A to B , and then to point C . Due to the symmetry of reflection the AB distance is the same

triangle:

$$tg\beta = \frac{EB}{d} \Rightarrow AF = EB = d tg\beta \quad (9.30)$$

$$AC = 2AF = 2d tg\beta \quad (9.31)$$

Substituting this back to equation (9.29) gives the optical path of the wave reflected back from the air/water interface:

$$l_2 = AD = AC \sin\alpha = 2d tg\beta \sin\alpha \quad (9.32)$$

Thus the optical path difference is:

$$\Delta l = l_1 - l_2 = \frac{2nd}{\cos\beta} - 2d tg\beta \sin\alpha \quad (9.33)$$

$$\Delta l = 2nd \left(\frac{1}{\cos\beta} - \frac{tg\beta \sin\alpha}{n} \right) \quad (9.34)$$

According to Snell's law:

$$1 \sin\alpha = n \sin\beta \Rightarrow \frac{\sin\alpha}{n} = \sin\beta \quad (9.35)$$

Substituting this to (9.34) gives

$$\Delta l = 2nd \left(\frac{1}{\cos\beta} - tg\beta \sin\beta \right) \quad (9.36)$$

$$\Delta l = 2nd \frac{1 - \sin^2\beta}{\cos\beta} \quad (9.37)$$

According to the Pythagoras theorem:

$$\sin^2\beta + \cos^2\beta = 1 \Rightarrow 1 - \sin^2\beta = \cos^2\beta \quad (9.38)$$

Thus:

$$\Delta l = 2nd \frac{\cos^2\beta}{\cos\beta} = 2nd \cos\beta \quad (9.39)$$

Each wavelength optical path difference is responsible to a 2π phase difference, therefore the phase difference due to the different optical path lengths is:

$$\Delta\phi_l = 2\pi \frac{\Delta l}{\lambda} = 2\pi \frac{2nd \cos\beta}{\lambda} \quad (9.40)$$

But there is one more thing to consider. It is a known fact, that when light is reflected back from the surface of an optically denser medium, it suffers a π phase shift. Therefore the wave which is reflected back from the air/water interface suffers a π phase shift. But the other wave, which is reflected back from the water/air interface does not suffer such a phase shift. This causes an extra π phase difference between the two waves in top of the phase difference due to their different optical path lengths.

$$\Delta\phi_r = \pi \quad (9.41)$$

$$\Delta\phi = \Delta\phi_l + \Delta\phi_r = 2\pi \frac{2nd\cos\beta}{\lambda} + \pi \quad (9.42)$$

When the phase difference is 0, or 2π , or 4π , etc... the two waves meet in phase. Therefore the criterion of constructive interference is:

$$2\pi \frac{2nd\cos\beta}{\lambda} + \pi = 2m\pi \quad (9.43)$$

$$2nd\cos\beta = \left(m - \frac{1}{2}\right) \lambda \quad (9.44)$$

where m is an integer number. The criterion of destructive interference is:

$$2\pi \frac{2nd\cos\beta}{\lambda} + \pi = (2m + 1)\pi \quad (9.45)$$

$$2nd\cos\beta = m\lambda \quad (9.46)$$

The criterion for constructive and destructive interference may be determined in a similar fashion for transmission too. (See figure 9.12) In this case while one wave travels on the ABH path, the other on the $ABCG$ path. Their optical path difference is again:

$$\Delta l' = 2nd\cos\beta \quad (9.47)$$

The only difference is that in this case neither wave suffers a π phase shift at reflection. Thus the criterion of constructive interference is:

$$2nd\cos\beta = m\lambda \quad (9.48)$$

While the criterion for destructive interference is:

$$2nd\cos\beta = \left(m - \frac{1}{2}\right) \lambda \quad (9.49)$$

Note, that the criteria for constructive and destructive interference for reflection and for transmission are precisely the opposite of each other. When the reflection is maximal, the transmission is minimal, and vice versa. Just like in case of diffraction, interference

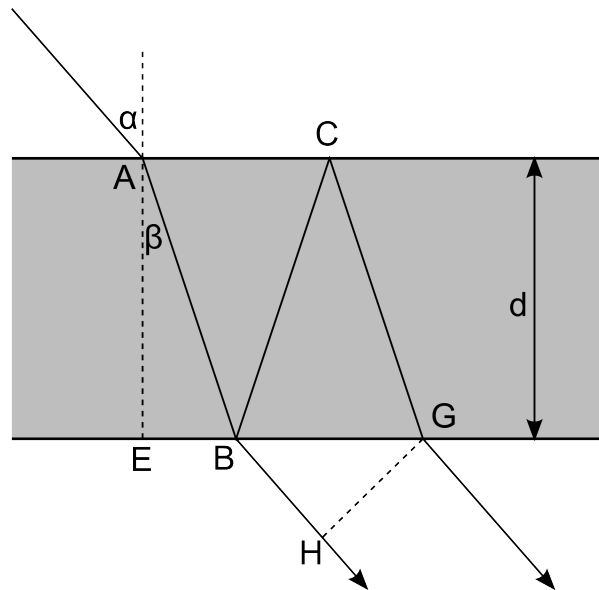


Figure 9.12: Transmission through a thin layer

cannot create or destroy energy: a destructive interference does not mean that the energy of light is lost. It is simply directed somewhere else.

It must also be noted, that the above deduced criteria would be the same if the thin layer would have lower index of refraction than its surroundings. (Such as an air gap between two pieces of glass) Just like above, one of the waves would suffer a π phase shift at reflection. The only difference is that in that case it would be the wave which travels through the layer.

On the other hand, if the thin layer is sandwiched between a medium with lower index of refraction, and another one with higher index of refraction, the criteria for constructive and destructive interference would be reversed, since in that case either both, or neither of the waves would suffer a π phase shift at reflection.

Thin layer interference occurs not only on oil spills and soap bubbles: it also has widespread applications in optics. As we have seen above, a thin layer may reduce reflection from a surface. Of course a single layer reduces reflection only at wavelengths that satisfy equation (9.46) or (9.49), depending on the index of refraction of the substrate onto which the layer has been deposited. (Don't forget, that if the index of refraction of the substrate is higher than that of the layer, the criteria are reversed.) But it is also possible to design layer structures consisting of a large number of thin layers with carefully chosen thicknesses and refractive indices that can reduce reflection in a wide band of wavelength. Such anti reflection coatings are usually deposited to the surfaces of different optical components. (For example in case of photography the so called lens flare effect is caused by reflections on the surfaces of lenses inside the objective of the

camera. By depositing a carefully designed anti reflection coating to these surfaces, the lens flare effect may be greatly reduced.) Similar layers are deposited to prescription glasses, too.

Laser mirrors also utilise thin layer interference. Traditional mirrors are usually produced by depositing a metal layer on a glass substrate. The problem is that the metal will always adsorb a part of the incident light, which may reduce the efficiency of the optical system and also damage the mirror itself. But using thin layer interference it is possible to design a layer structure made entirely of dielectric materials that can effectively reflect the laser light without adsorbing any of it.

Carefully designed layer structures may be tailored to reflect or transmit certain wavelengths or wavelength bands, while blocking others. These are usually referred to as interference filters.

Thin layer interference may even be used for decorative purposes. When exposed to air, most metals starts oxidising, until a stable oxide layer forms on their surface and protects them from further oxidation. The thickness of this so called native oxide layer is usually very small (sometimes less than a nanometre), but it can be increased relatively simply by an electrochemical processes (the so called anodic oxidation) or a heat treatment in an oxidising atmosphere. Just like the colour reflected back from a soap bubble depends on the thickness of its wall, so will the colour of the metal depend on the thickness of its oxide layer. Certain metals (such as aluminium) can be relatively easily coloured this way, without the use of any paint. Since aluminium oxide is very hard and very stable so called eloxed aluminium will be more durable, and maintain its colour much longer, than any paint.

Chapter 10

Einstein's Special Theory of Relativity – Gábor Dobos

10.1 The Aether Hypothesis and The Michelson-Morley Experiment

Since Young's diffraction experiments in the early 19th century scientists knew that light is a wave. But all other types of waves required some sort of medium to exist. For example a ringing alarm clock may be silenced by placing it into a vacuum chamber, and pumping the air out. Without the presence of air sound waves cannot reach us, therefore we won't hear the alarm. But if there is a window on the chamber we will still see the clock. This means that light waves can travel through vacuum. 19th century scientists were left with a tough question: what is the medium through which light waves are travelling? What remains in the vacuum chamber, after air was pumped out? What is the medium filling the vacuum of space, and allowing sunlight and the light of distant stars to reach us?

To answer the question scientists have postulated the existence of a medium called luminiferous (or lightbearing) aether. It must be noted however that even 19th century scientists have realised that this hypothetic medium should have rather peculiar properties. Aether must be omnipresent, as light is capable to propagate through vacuum. On the other hand it cannot be solid, since objects can move through it without any measurable resistance. This raises another difficult problem: certain physical phenomena (such as birefringence) cannot be explained by longitudinal waves, only by transverse waves. As we have already mentioned in the previous semester transverse waves can exist only in solids. But if aether is a solid, how could other solid objects (such as planets and other celestial bodies) pass through it without any resistance?¹

As an attempt to resolve the contradiction Augustin Louis Cauchy suggested that

¹Thomas Young was not the first scientist who attempted to describe light as a wave. Christiaan

aether may behave like a non-Newtonian fluid. The viscosity of certain materials (such as a suspension of corn starch in water) depends on the shear rate: when we try to change their shape slowly they flow like a fluid, but they resist quick deformations as if they were solid. Since the frequency of light waves is very high, it seemed to be possible, that aether may act as a solid at these high frequencies (thus enabling transverse waves), while flow virtually without any resistance at low frequencies. (It must be noted however, that it seemed to be hard to imagine a substance which has practically zero viscosity at low frequencies, yet it proves to be several orders of magnitude stiffer than any other material at high frequencies...) A further problem was that aether had to be massless, otherwise its gravity would influence the orbit of planets.

The aether hypothesis persisted even after Maxwell had given a full description of electrodynamics, which led to the realisation that light is a form of electromagnetic waves. These are basically oscillating electric and magnetic fields, and scientists believed that only separated positive and negative electrical charges may create such dipolar electric fields. Since electrical charge is an inextricable property of matter, this also suggested that light must propagate through some sort of medium.²

According to Maxwell's equations electromagnetic waves travel at a constant velocity, which depends only on the magnetic permeability and electric permittivity of the medium. Since scientists believed that aether is the medium which carries electromagnetic waves, it seemed to be obvious that this velocity has to be measured with respect to the aether. This also meant that it could serve as a universal frame of reference: it should be possible to determine the velocity of any observer with respect to the omnipresent aether by measuring the speed of light. (If light is propagating through aether at a constant velocity, an observer moving with respect to aether should find that the speed of light in his frame of reference is slightly altered by his own velocity with respect to aether.) The problem was that even the velocities of celestial objects are small compared to the speed of light. Thus instruments had to be very precise to be able to detect such small variations. The first apparatus that had the required precision was constructed by Michelson and Morley in 1887.

The idea was that the Earth is moving at a relatively high velocity around the Sun. (Its average orbital speed is 29.78 km/s , which is only four orders of magnitude smaller than the speed of light.) As the Earth orbits the Sun the direction and magnitude of its velocity is changing continuously. This means that even if it happens to be stationary with respect to aether at a given day when we perform the experiment, this shall not last: as it continues to orbit the sun, its velocity has to change, and it cannot remain stationary. This means, that the speed of light should be a few km/s different in the

Huygens has already suggested this interpretation in the 17th century, but it was rejected by Newton because of the above described contradiction.

²Although electromagnetic induction of electric fields in vacuum is not forbidden by Maxwell's equations it cannot be directly demonstrated since all methods of detecting electric fields require the presence of electrically charged particles.

direction of Earth's movement than in the perpendicular direction, and this difference (and the direction in which it is detected) should change annually. To detect this difference Michelson constructed a device capable of comparing the speed of light in two perpendicular directions.

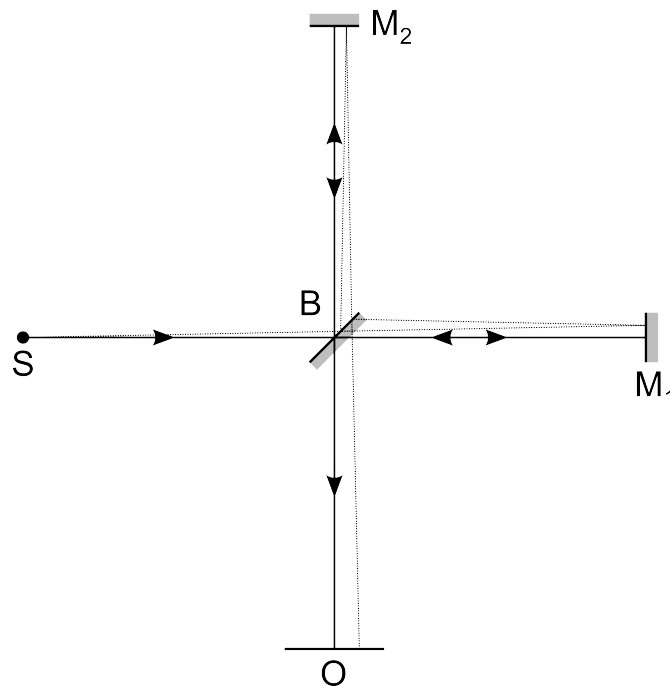


Figure 10.1: A simplified Michelson interferometer

A simplified sketch of a Michelson interferometer is shown on figure 10.3. Light coming from the source (S) is split into two beams by a half-silvered mirror B . (This is usually referred to as beam splitter.) Some of the light is transmitted towards mirror M_1 , while the rest is reflected back towards M_2 . (The BM_1B and BM_2B paths, where the two beams travel separately are usually referred to as the arms of the interferometer.) After both beams are reflected back from their respective mirrors they are recombined once again at the beam splitter: some of the light coming from M_1 is reflected back towards the observer (O), while some of the light reflected back from M_2 is transmitted by the beam splitter in the same direction. These two beams interfere with each other, and the intensity detected by the observer³ depends on their phase difference: it is maximal

³In practice the interference is usually observed on a screen. This makes it possible to see the interference of slightly off-axis rays (dotted line of figure 10.3). These rays have to travel a slightly longer distance, which proportionally increases the phase difference between the interfering waves. The result is that the intensity on the screen varies periodically with the distance from the centre, forming a series of bright and dark rings that are called diffraction fringes. (It is also possible to produce a line-pattern on the screen instead of circles by slightly tilting both mirrors.)

when they meet in phase, and minimal when they meet in opposing phases. The phase difference in turn depends on the time light requires to travel along each arm of the interferometer. This means, that if the length of the arms are adjusted to be equal, the interferometer may be used to detect differences in the speed of light in the direction of its two arms.

The precision of the interferometer depends on the length of its arms. The longer the arms, the larger the phase difference is going to be due to the same difference in the speed of light. To improve precision Michelson and Morley constructed an interferometer, where light was reflected back and forth several times along the arms, increasing the path to 11 metres. To reduce vibrations and temperature variation, the experiment was performed in a closed basement room.

To further decrease vibrations the interferometer itself was assembled on a large sandstone block, which was floating in a pool of mercury. This also made it easy to turn the entire device. After a single push it would keep rotating for a relatively long time, making it possible to scan deviations in the speed of light in different directions without touching the interferometer. (Touching the device would create so much vibration, that it would make the measurement impossible...) Before the start of the experiment the interferometer was adjusted until the phase difference between the beams was zero, then it was given a small push, to start its rotation. This made it possible to observe how the diffraction pattern changed as the device was slowly rotating. When one of its arms pointed in the direction of the movement of Earth the speed of light should have been altered in that direction. This could be detected as a shift in the diffraction pattern. As the device was rotating, the arms would point alternately in the direction of Earth's movement, or in the perpendicular direction, periodically shifting the diffraction pattern.

Michelson and Morley estimated that they might be able to detect if the pattern was shifted by one hundredth of a diffraction fringe. At the same time the orbital velocity of the Earth around the Sun should result in a 0.4 fringe shift, which should be well detectable. But contrary to their expectations they have found the speed of light to be the same in all directions.⁴

⁴It must be noted, that there could be another interpretation. If we try to measure the speed of sound, it proves to be the same in all directions, and show no annual variations, despite of Earth's movement around the Sun. The reason is that our planet is carrying its own atmosphere with it, thus the velocity of Earth with respect to the air surrounding it is practically zero. It seemed to be possible that objects may drag aether with them, the same way as planets are carrying their atmosphere. This was referred to as the aether drag hypothesis. But complete aether drag is not compatible with the results of certain experiments (such as the Fizeau experiment) and astronomical observations (such as stellar parallax measurements) that were performed decades before the Michelson-Morley experiment, therefore this cannot be a valid interpretation of the results.

10.2 Einstein's Special Theory of Relativity

The result of these experiments proved to be very hard to explain. According to classical physics the movement of the observer should influence the speed of light, but experiments have proven this expectation to be wrong. It was Einstein's special theory of relativity that gave a proper interpretation of these results in 1905. The principle of relativity was well known even in classical mechanics. It stated that all inertial frames of reference are equivalent: mechanical laws are the same in all inertial frames. For example an observer sitting on a train whose curtains are drawn cannot perform any mechanical experiments that could determine if the train is moving or not.

But this principle would have been violated by the positive result of the Michelson-Morley experiment. It would have created a means to measure the velocity of an observer with respect to aether. Thus not all inertial frames of reference would be equivalent: there would be a special, privileged frame (the one which is at rest with respect to the aether) to which all velocities should be measured. The principle of relativity would be true for mechanics (as there are no mechanical experiments, through which inertial frames could be distinguished), but it would not apply to other disciplines of physics, such as electrodynamics. Einstein believed that this is unacceptable: there is only one physical world, and it is guided by one set of rules. There should be a symmetry to the laws of nature: the most basic principles (such as relativity) should apply to all fields of science.

Einstein's idea was surprisingly simple and elegant: let us accept relativity as a general principle that applies to all of nature, not only mechanics. And let us also accept the result of the Michelson-Morley experiment, which shows that the movement of the observer does not influence the speed of light. Einstein summarised his assumptions in two postulates:

1. All inertial observers are equivalent with respect to ALL natural phenomena. There are no special, privileged frames of reference.
2. The speed of light is the same in all inertial frames, irrespectively of the state of movement of the lightsource or the observer.

To understand the consequences of these assumptions let us follow a simple thought-experiment. Let us consider two inertial frames of reference K and K' . Let the x axis of both coordinate systems point in the same direction, and let K' move in this direction at a constant u velocity with respect to K . For the sake of simplicity let us synchronise the clocks of two observers sitting in the origins of K and K' when they pass each other. (Thus $t_0 = t'_0 = 0$ when $x_0 = x'_0 = 0$.) At the moment when the two origins coincide let us send a light pulse from that point to the positive x direction, and let us try to measure its velocity in both frames of reference. Velocity measurements are relatively straightforward: all we have to do is measure the time light requires to travel a given distance, and then calculate the ratio of the distance and the travel time. Again, for the

sake of simplicity let us pick a point on the x axis, and instruct the observers in both K and K' to measure the time, light requires to reach this point. The observer in the K frame will find, that the light pulse reaches this point at a t moment, and its position is x . Let us mark the time measured by the observer in K' by t' , and the coordinate of the point by x' .

According to classical mechanics there is a connection between the values measured in K and K' , which is called Galilean transformation:

$$x' = x - ut \quad (10.1)$$

$$t' = t \quad (10.2)$$

Or:

$$x = x' + ut' \quad (10.3)$$

From this, the speed of light measured in K' should be:

$$c' = \frac{x'}{t'} = \frac{x - ut}{t} = \frac{x}{t} - u = c - u \quad (10.4)$$

where c is the speed of light in K , and c' is the speed of light in K' . This means that if the Galilean transformation would be true, the speed of light should be different in different frames of reference, which would contradict Einstein's second postulate and the results of the experiments discussed in the previous section.

This contradiction clearly shows that there is something wrong with the Galilean transformation. If we carry out the experiment in practice, both observers will measure the speed of light to be the same. From the viewpoint of an observer in K the fact that the other observer in K' gets the same result to the speed of light would seem like he made some mistakes during his measurements. It would seem like the observer in K' has measured either the distance or the time (or both) incorrectly, as if his instruments were not calibrated correctly. (Imagine, that u is three quarters of the speed of light. In this case, the observer in K would expect that the observer in K' will measure the speed of light to be one quarter light speed. Instead, he reports that he have measured the speed of light to be the same as in K . From the viewpoint of the observer in K , this would seem as if the results of the other observer are off by a factor of four.) The observer in K may take this into account by multiplying every value measured by the observer in K' by a constant factor:

$$x' = \gamma(x - ut) \quad (10.5)$$

But the observer in K' would believe the same about the observer in K . He would think, that his own measurements are correct, and the measurements of the observer in K are the ones that are distorted. Just like his colleague in K , he may try to correct this

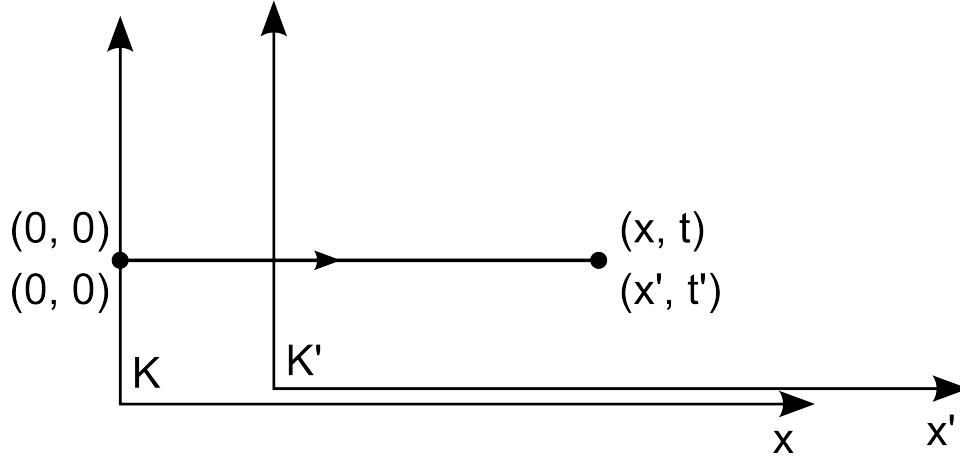


Figure 10.2: Observers in both K and K' measures the speed of light to be the same ($c = \frac{x}{t} = \frac{x'}{t'}$). This contradicts the predictions of classical physics, thus the Galilean transformation has to be modified.

by multiplying every measurement made by the observer in K by a constant factor. But remember: the principle of relativity dictates that all inertial observers are equivalent. If one is thinking about the other that his measurements are off by a factor of γ , the other should think the same about him. Thus the observer in K' should multiply the values measured by the observer in K by the same factor:

$$x = \gamma(x' + ut') \quad (10.6)$$

We may determine the value of γ from equation (10.5) and (10.6). Let us multiply the left hand side of (10.5) by the left hand side of (10.6) and the right hand side of (10.5) by the right hand side of (10.6):

$$xx' = \gamma^2(x' + ut')(x - ut) \quad (10.7)$$

$$\gamma = \sqrt{\frac{xx'}{(x' + ut')(x - ut)}} \quad (10.8)$$

$$\gamma = \sqrt{\frac{xx'}{(xx' + ut'x - utx' - u^2tt')}} \quad (10.9)$$

$$\gamma = \frac{1}{\sqrt{(1 + u\frac{t'}{x'} - u\frac{t}{x} - u^2\frac{t}{x}\frac{t'}{x'})}} \quad (10.10)$$

According to Einstein's second postulate the speed of light is the same in both frames of

reference, thus $\frac{t'}{x'} = \frac{t'}{x'} = \frac{1}{c}$. Substituting this to the equation above gives:

$$\gamma = \frac{1}{\sqrt{(1 + uc - uc - \frac{u^2}{c^2})}} \quad (10.11)$$

$$\gamma = \frac{1}{\sqrt{(1 - \frac{u^2}{c^2})}} \quad (10.12)$$

From this:

$$x' = \frac{x - ut}{\sqrt{(1 - \frac{u^2}{c^2})}} \quad (10.13)$$

Using this, t' may be determined from (10.6):

$$t' = \frac{t - \frac{ux}{c^2}}{\sqrt{(1 - \frac{u^2}{c^2})}} \quad (10.14)$$

Equations (10.13) and (10.14) are called Lorentz transformation and they replace the Galilean transformation in Einstein's special theory of relativity. Unlike Galilean transformation, Lorentz transformation ensures that the speed of light is the same in all inertial frames of reference. This means that we shall abandon Galilean transformation, and use Lorentz transformation instead, because it is the one, which complies with the results of our observations. It has to be noted however that this change has some very interesting consequences, as we will see in the following sections.

10.3 Lorentz contraction and time dilatation

In classical physics (using Galilean transformation) the length of an object was the same in all frames of reference. In Einstein's theory of relativity this is no longer true. Imagine that we have a rod that is at rest in the K' frame (which is moving at a constant u velocity with respect to the K frame in the positive x direction). Let us mark the length of the rod in the K frame by $L = x_2 - x_1$, and by $L' = x'_2 - x'_1$ in the K' frame. According to

the Lorentz transformation the length of the rod in the K' frame is:

$$L' = x'_2 - x'_1 \quad (10.15)$$

$$L' = \frac{x_2 - ut}{\sqrt{1 - \frac{u^2}{c^2}}} - \frac{x_1 - ut}{\sqrt{1 - \frac{u^2}{c^2}}} \quad (10.16)$$

$$L' = \frac{x_2 - x_1}{\sqrt{1 - \frac{u^2}{c^2}}} \quad (10.17)$$

$$L' = \frac{L}{\sqrt{1 - \frac{u^2}{c^2}}} \quad (10.18)$$

Or:

$$L = L' \sqrt{1 - \frac{u^2}{c^2}} \quad (10.19)$$

This means that from the viewpoint of an observer in the K frame, the length of the moving rod has been reduced by a factor of $\sqrt{1 - \frac{u^2}{c^2}}$. According to Einstein's theory of relativity moving objects contract in the direction of their movement with respect to their rest length. (The length of the rod in the co-moving frame is referred to as rest-length.) This phenomenon is called Lorentz contraction.

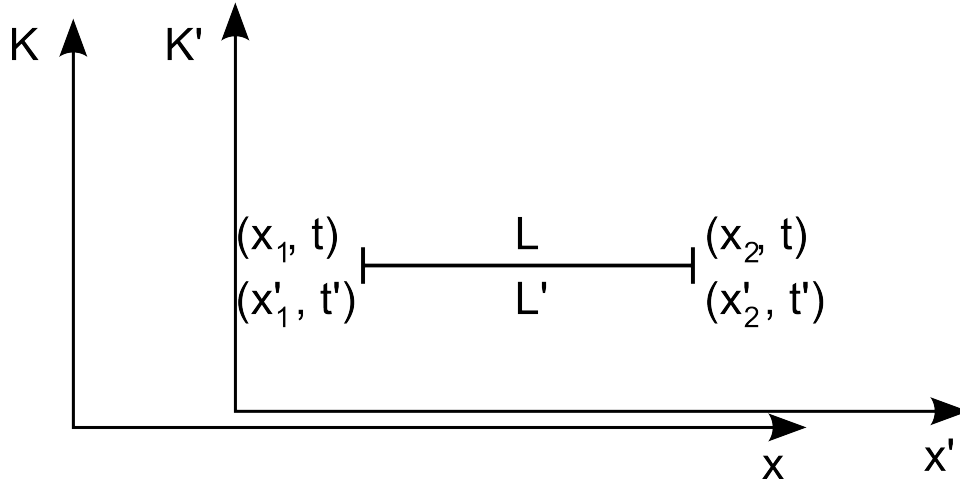


Figure 10.3: Due to Lorentz contraction moving objects contract in the direction of their movement with respect to their rest length

In classical physics time was the same in all frames of reference. According to equation (10.14) this is no longer true in relativistic physics. To understand the phenomena,

imagine that we try to measure the time that has passed between two events in both the K and K' frames. (For example, imagine that a really fast race car is driving at a constant velocity between the start and finish line of a straight race track. We try to measure the time the car required to drive along the track. The two events are the car passing the start and the finish line.) Again, the K' frame (the race car) is moving at a constant u velocity in the positive x direction with respect to the K frame (the race track). For the sake of simplicity let the two events take place at the same location in the K' frame, thus $\Delta x' = x'_2 - x'_1 = 0$. (The driver's stopwatch is moving together with the race car, thus its position with respect to the race car does not change.) Since the K' frame is moving with respect to the K frame, the two events do not take place at the same location in the K frame. (The first event takes place at the start line, while the second event takes place at the finish line. The distance between the positions of the two events in the racetrack's frame of reference is the length of the track.) Let us mark the time between the two events in the K frame by $\Delta t = t_2 - t_1$ (This is the time measured by the racetrack's built-in timing system.), and in the K' frame by $\Delta t' = t'_2 - t'_1$. According to the Lorentz transformation:

$$\Delta t = t_2 - t_1 \quad (10.20)$$

$$\Delta t = \frac{t'_2 + \frac{ux'_2}{c^2}}{\sqrt{1 - \frac{u^2}{c^2}}} - \frac{t'_1 + \frac{ux'_1}{c^2}}{\sqrt{1 - \frac{u^2}{c^2}}} \quad (10.21)$$

$$\Delta t = \frac{t'_2 - t'_1 + \frac{u(x'_2 - x'_1)}{c^2}}{\sqrt{1 - \frac{u^2}{c^2}}} \quad (10.22)$$

$$\Delta t = \frac{t'_2 - t'_1}{\sqrt{1 - \frac{u^2}{c^2}}} \quad (10.23)$$

$$\Delta t = \frac{\Delta t'}{\sqrt{1 - \frac{u^2}{c^2}}} \quad (10.24)$$

Since $\sqrt{1 - \frac{u^2}{c^2}}$ is always smaller than 1, Δt is always longer than $\Delta t'$. (The official timing system of the racetrack will always measure the run-time of the driver to be slightly longer than his own measurement. From the viewpoint of the observers standing along the racetrack, the driver's clock is not accurate: it is ticking slightly slower than the stationary clock of the race track, as if time has slowed down in the race car.) The phenomenon is referred to as time dilatation.

To better understand the relationship between Lorentz contraction and time dilatation imagine that mankind is planning its first interstellar expedition because our astronomers have found a habitable planet 200 light years away from Earth. The problem is that no space ships can travel faster than the speed of light, and since we have no working hibernation technology one would think, that none of the explorers would live long enough to reach their destination. But this is not the case. The theory of relativity offers a solution: if we can build a space ship, that is fast enough, Lorentz contraction and time dilatation would make the journey feasible.

Let's say, that we are able to build a space ship that can travel at 99% of the speed of light. In this case the value of the $\sqrt{1 - \frac{u^2}{c^2}}$ factor is approximately 0.141. Although it takes the spaceship 202 years to travel 200 light years at this velocity, the astronauts would still survive the trip. From the viewpoint of the people left behind on earth, time dilatation slows down time aboard the spaceship. Even though an earthbound observer would say that the journey took 202 years, the explorers aboard the spaceship would age only 28.5 years.

But the astronauts on the spaceship would tell a different story. Remember that all motion is relative! From their viewpoint the entire universe is moving with respect to them at 99% of the speed of light. You should also remember that all moving bodies contract in the direction of their motion due to the Lorentz contraction. This means that from the viewpoint of the astronauts the distance they have to travel has shrunk to 28.2 light years, and traveling at 99% light speed they can pass this distance in 28.5 years.

Notice that although the stories of the earthbound observers and the astronauts are very different, the end result is the same: the astronauts have reached their destination, and they aged only 28.5 years during the journey. From the viewpoint of the earthbound observers this was due to time dilatation, while the astronauts would claim that it was Lorentz contraction which made their journey possible.

Although this thought experiment may seem to be very futuristic, and it is unlikely that we will organise such an expedition in the foreseeable future, these effects have been verified by experiments. We might not be able to accelerate a large space ship to 99% of the speed of light, but we are capable to accelerate elementary particles to such high velocities. Just like people, certain elementary particles have limited lifetimes: they decay to other particles. According to classical physics this should limit the distance they can travel from their point of origin at a given velocity. But experiments show that (just like our fictional astronauts) they may reach considerably larger distances, due to relativistic effects. The results of these experiments show a good agreement with the predictions of Einstein's theory of relativity.

The important lesson to remember is that space and time are not absolute: they depend on your point of view. Different observers may give different descriptions of the same experiment, but the end result is always the same. There is only one physical reality, and although the interpretation of events may vary with the viewpoint of the

observer, the laws of nature are the same in all inertial frames of reference.

10.4 Velocity addition

Imagine that there is an object moving at a v velocity in the positive x direction in the K inertial frame. Let us try to calculate the velocity of this object in the K' frame, which is (yet again) moving at a constant u velocity in the same direction with respect to K . As we have noted earlier velocity measurements are really straightforward: all we have to do is calculate the distance the object travels in a given time, and calculate the ratio of the two values. If in the K' frame the object is at the x'_1 position at the t'_1 moment, and at the x'_2 position at the t'_2 moment, its velocity is:

$$v' = \frac{x'_2 - x'_1}{t'_2 - t'_1} \quad (10.25)$$

We may use equations (10.13) and (10.14) to determine x'_1 , x'_2 , t'_1 and t'_2 from the corresponding coordinates in K :

$$v' = \frac{\frac{x_2 - ut_2}{\sqrt{1 - \frac{u^2}{c^2}}} - \frac{x_1 - ut_1}{\sqrt{1 - \frac{u^2}{c^2}}}}{\frac{t_2 - \frac{ux_2}{c^2}}{\sqrt{1 - \frac{u^2}{c^2}}} - \frac{t_1 - \frac{ux_1}{c^2}}{\sqrt{1 - \frac{u^2}{c^2}}}} \quad (10.26)$$

$$v' = \frac{x_2 - ut_2 - (x_1 - ut_1)}{t_2 - \frac{ux_2}{c^2} - (t_1 - \frac{ux_1}{c^2})} \quad (10.27)$$

$$v' = \frac{\Delta x - u\Delta t}{\Delta t - \frac{u}{c^2}\Delta x} \quad (10.28)$$

$$v' = \frac{v - u}{1 - \frac{uv}{c^2}} \quad (10.29)$$

$$(10.30)$$

We may also rearrange (10.30) to calculate the velocity of the object in K from the velocity in K'

$$v = \frac{v' + u}{1 + \frac{uv'}{c^2}} \quad (10.31)$$

Imagine that an object is moving at $0.75c$ in the K' frame, and K' is also moving at $0.75c$ with respect to K . In classical physics, the Galilean transformation would predict, that an observer in the K frame would find the velocity of the object to be $1.5c$. But according to relativistic calculations the velocity of the object in the K frame is only

$$v = \frac{0.75c + 0.75c}{1 + 0.5625} = 0.96c \quad (10.32)$$

which is smaller than the speed of light. Even if u and v' is 99% of the speed of light, v is going to be only $0.99995c$. It doesn't matter how close u and v' comes to the speed of light, the velocity of the object can never be higher than the speed of light in any frame of reference. This shows that the speed of light is not only the same in all frames of reference, it is also maximal velocity of any object.

10.5 Connection between relativistic and classical physics

Throughout this chapter we have discussed the differences between relativistic physics, and classical physics. The formulas we have to use in relativistic calculations are markedly different from the formulas of classical mechanics, and there are a series of phenomena that could never exist in classical physics. It may seem like there is a direct contradiction between classical and relativistic physics.

But how is this possible? The scientific method requires that all theories must be based on measurements and observations, and the predictions of the theories should be tested by further experiments. The laws of classical physics went through this process. They have been tested by countless measurements, and scientists haven't found any deviations from their predictions for centuries. If classical physics is wrong, how could it pass all these checks? And even more importantly: how can we trust any physical theories after this?

The answer is that classical physics is not really wrong, only less precise than modern physics. All scientific theories are based on experiments and measurements. The problem is that it doesn't matter how careful we are, no practical measurement can be completely precise. There is always some measurement error.⁵ The amount of measurement data also has practical limits. This means that theories are based on a finite number of measurements performed in a finite range of parameters with finite precision. If a theory is so close to the truth that the deviation is smaller than the error of our measurements, the problem may not be detected without new, more precise measurement techniques.

⁵The only exception is, if we are measuring a discrete quantity, for example the number of some objects. But even in such cases, absolute precision might not be guaranteed. For example the number of particles released by a nuclear process may be counted precisely, but such processes tend to be stochastic. This means that there is a random variation in the number particles released by the sample. The measurement data will show a statistical fluctuation, not because of the inaccuracy of our measurements, but because of the random nature of the process itself.

For example, classical mechanics was established based on measurements and observations performed at relatively low velocities (at least compared to the speed of light). At such velocities the predictions of relativistic physics are so close to the predictions of classical physics, that the deviation is hardly detectable, even with today's modern instruments. Consider the Lorentz transformation (equations (10.13) and (10.14)):

$$x' = \frac{x - ut}{\sqrt{1 - \frac{u^2}{c^2}}} \quad (10.33)$$

$$t' = \frac{t - \frac{ux}{c^2}}{\sqrt{1 - \frac{u^2}{c^2}}} \quad (10.34)$$

Although the formulas seem to be very different from the formulas of Galilean transformation (equations (10.1) and (10.2)), the deviation is very small at low velocities. The speed of light is three or four orders of magnitude higher than even celestial velocities. Therefore the $\frac{u^2}{c^2}$ term is usually a very small number. Thus $1 - \frac{u^2}{c^2}$ is very close to 1, and its square root is even closer. This means that unless the velocities involved are close to the speed of light, the denominator on the right hand side of both equations may be omitted. In a similar manner, it is easy to see that $\frac{ux}{c^2} \ll t$. Using these approximations:

$$x' = \frac{x - ut}{\sqrt{1 - \frac{u^2}{c^2}}} \longrightarrow x - ut \quad (10.35)$$

$$t' = \frac{t - \frac{ux}{c^2}}{\sqrt{1 - \frac{u^2}{c^2}}} \longrightarrow t \quad (10.36)$$

The Galilean transformation is the low velocity limit of the Lorentz transformation. In other words the predictions of the classical and the relativistic theory are practically the same at low velocities; no significant deviations can be detected. This is the reason why scientists haven't realised that there is something wrong with classical mechanics until the late 19th century. This also means that we may continue to use classical mechanics at these low velocities with impunity, despite of the fact, that we know that there are fundamental problems with the theory, because we also know that these problems would realise themselves only at very high velocities.

This also reveals the relationship between the laws of physics that you may find in textbooks, and the actual laws of nature. Scientists are trying to create a mathematical

model of the physical world around us. No one can guarantee that this model is completely precise, but scientists are always using the best and most precise techniques that are available at a given time to look for deviations. If any deviation is found the model is revised to account for the newly found phenomena.

This continuous, self-critical revision and improvement is the key to the success of science. We may trust that the predictions of scientific theories are always precise enough for any practical application, because they are always based on the best and most precise measurements that our civilisation can produce at a given time.