The Karlsruhe Physics Course

for the lower secondary school

Volume 2

Note to the reader

We have chosen a *one-section-one-page* layout. The advantage is that figures, tables and equations stay where they are supposed to be. Moreover, it is easy to update the text. For reading we recommend the Adobe reader or the GoodReader.

Friedrich Herrmann

The Karlsruhe Physics Course

A Physics Text Book for the Lower Secondary School
Volume 2

With the collaboration of Karen Haas, Matthias Laukenmann, Lorenzo Mingirulli, Petra Morawietz, Dieter Plappert and Peter Schmälzle

Illustrations: Friedrich Herrmann

Translation: GETS (German English Translation Services)
Robin Fuchs, Winterthur

TABLE OF CONTENTS

15 Data and Data Carriers

15.1 Data transfer
15.2 The amount of data
15.3 Examples of a data transport
15.4 Data currents

16 Electricity and Electric Currents

16.1 The electric circuit
16.2 Electric current
16.3 The junction rule
16.4 Electric potential
16.5 The zero point of electric potential
16.6 Driving force and currents
16.7 Applications
16.8 Electric resistance
16.9 Short circuits and fuses
16.10 Alternating current
16.11 The dangers of electric currents

17 Electricity and Energy

17.1 Electricity as an energy carrier
17.2 Transmission resistance – energy loss in wires

18 The Magnetic Field

18.1 Some simple experiments with magnets and nails
18.2 Magnetic poles
18.3 Lines of magnetization
18.4 Magnetic fields
18.5 Graphic representation of magnetic fields
18.6 Magnetization lines and field lines
18.7 Magnetic fields and matter
18.8 Energy of a magnetic field
18.9 Electric currents and magnetic fields
18.10 Electromagnets
18.11 Electric motors
18.12 The Earth's magnetic field
18.13 Induction
18.14 Generators
18.15 Transformers
18.16 The magnetic field of induced currents
18.17 Superconductors

19 Electrostatics

19.1 Charge and charge carriers
19.2 Charge currents and charge carrier currents
19.3 Accumulation of electric charge
19.4 Electric fields
19.5 Capacitors
19.6 Capacitance
19.7 Cathode ray tubes
19.8 Atmospheric electricity

20 Data Systems Technology

20.1 Amplifiers
20.2 Data processing
20.3 Generalizing the definition of the amount of data

21 Light

21.1 Light sources
21.2 Some characteristics of light
21.3 When light meets matter
21.4 Diffuse and coherent light
21.5 Reflection law
21.6 Plane mirrors
21.7 Parabolic mirrors
21.8 Refraction of light
21.9 Prisms
21.10 Total reflection

22 Optical Image Formation

22.1 What is an image?
22.2 Pinhole cameras
22.3 The relation between the object size and picture size
22.4 Improving the pinhole camera
22.5 Lenses
22.6 Making optical images with lenses
22.7 Focal distance and refractive power
22.8 Combining lenses
22.9 Depth of field
22.10 Objective lenses
22.11 Cameras
22.12 The eye
22.13 Glasses and magnifying glasses
22.14 Slide projectors an overhead projectors
22.15 Film cameras, film projectors, video cameras
22.16 Microscopes
22.17 Telescopes
22.18 Astronomical telescopes

23 Color

23.1 Three dimensional color space
23.2 Mixing light
23.3 How the eye can be deceived – television images
23.4 Back to color space
23.5 Spectra
23.6 The relationship between spectrum and color impression

15

Data and Data Carriers

15.1 Data transfer

Every house is connected to the outside world by cables, wires and openings. Figures 15.1 and 15.2 show a schematic of a house with these connections. In order not to lose the overview, one part of the connections is shown in the first Figure and another part in the second. They have been grouped according to a particular principle. All the connections sketched in Fig. 15.1 serve a common purpose, and all the ones represented in Fig. 15.2 serve another common purpose.

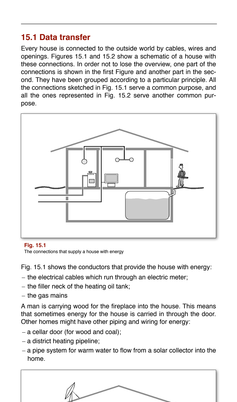

Fig. 15.1
The connections that supply a house with energy

Fig. 15.1 shows the conductors that provide the house with energy:
- the electrical cables which run through an electric meter;
- the filler neck of the heating oil tank;
- the gas mains

A man is carrying wood for the fireplace into the house. This means that sometimes energy for the house is carried in through the door. Other homes might have other piping and wiring for energy:
- a cellar door (for wood and coal);
- a district heating pipeline;
- a pipe system for warm water to flow from a solar collector into the home.

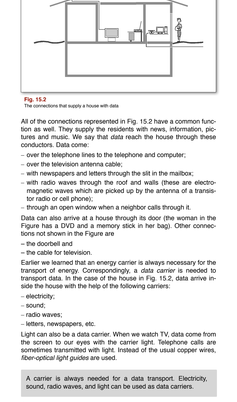

Fig. 15.2
The connections that supply a house with data

All of the connections represented in Fig. 15.2 have a common function as well. They supply the residents with news, information, pictures and music. We say that *data* reach the house through these conductors. Data come:
- over the telephone lines to the telephone and computer;
- over the television antenna cable;
- with newspapers and letters through the slit in the mailbox;
- with radio waves through the roof and walls (these are electromagnetic waves which are picked up by the antenna of a transistor radio or cell phone);
- through an open window when a neighbor calls through it.

Data can also arrive at a house through its door (the woman in the Figure has a DVD and a memory stick in her bag). Other connections not shown in the Figure are
- the doorbell and
- the cable for television.

Earlier we learned that an energy carrier is always necessary for the transport of energy. Correspondingly, a *data carrier* is needed to transport data. In the case of the house in Fig. 15.2, data arrive inside the house with the help of the following carriers:
- electricity;
- sound;
- radio waves;
- letters, newspapers, *etc.*

Light can also be a data carrier. When we watch TV, data come from the screen to our eyes with the carrier light. Telephone calls are sometimes transmitted with light. Instead of the usual copper wires, *fiber-optical light guides* are used.

> A carrier is always needed for a data transport. Electricity, sound, radio waves, and light can be used as data carriers.

Of course, any message can be transmitted by any of the carriers. In Fig. 15.3, the message "Write me an email" is transferred with four different carriers.

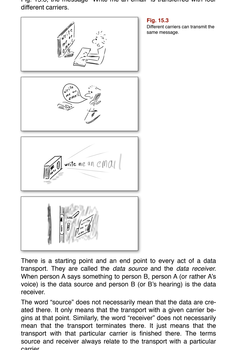

Fig. 15.3
Different carriers can transmit the same message.

The word "source" does not necessarily mean that the data are created there. It only means that the transport with a given carrier begins at that point. Similarly, the word "receiver" does not necessarily mean that the transport terminates there. It just means that the transport with that *particular* carrier is finished there. The terms source and receiver always relate to the transport with a particular carrier.

In this way, messages that are transmitted over a telephone cable come from the microphone of the telephone. This microphone is the data source for the transport through the cable (with the carrier electricity). The loudspeaker in the other telephone is the corresponding data receiver.

A data transport can be represented in a flow diagram in analogy to the transport of energy. Data source and data receiver are each symbolized by a box. The data flow is represented by a fat arrow, and the current of the data carrier by a thin one. Fig. 15.4 shows three examples of a data flow diagram.

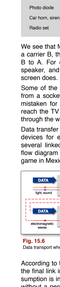

Fig. 15.4
Some data flow diagrams

There is a starting point and an end point to every act of a data transport. They are called the *data source* and the *data receiver.* When person A says something to person B, person A (or rather A's voice) is the data source and person B (or B's hearing) is the data receiver.

Table 15.1 shows further source-receiver pairs together with their corresponding carriers.

Table 15.1

Data source	Data carrier	Data receiver
Sender antenna	Electromagnetic waves	Receiver antenna
Receiver antenna	Electricity	Television set
Speech organs	Sound	Microphone

When energy is transported, we often encounter devices that transfer energy from one carrier to another. This is also the case in data transport. In data transport, devices are often involved which transfer data from one carrier to another. This means that such devices are both source and receiver.

This is the case when data are transferred from electricity to sound in a loudspeaker. The loud speaker is a receiver for transport with the carrier electricity and a source for transport with the carrier sound. The symbolic representation of a data transfer device is obvious. Fig. 15.5 shows some examples.

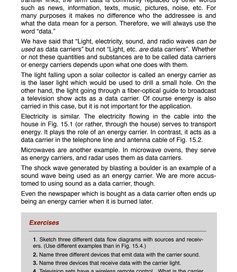

Fig. 15.5
Data transfer units

In table 15.2, more data transfer devices are shown along with their corresponding carriers at inlets and outlets.

Table 15.2

	Data carrier	
Data transfer unit	**at the inlet**	**at the outlet**
Liquid crystal display	Electricity	Light
Photo diode	Light	Electricity
Car horn, siren	Electricity	Sound
Radio set	Electromagnetic waves	Sound

We see that for every device which transfers data from a carrier A to a carrier B, there is a corresponding device that transfers data from B to A. For example, the microphone is the analog to the loudspeaker, and the video camera does the opposite of what a TV screen does.

Some of the devices in Fig. 15.5 need to be supplied with energy from a socket in order to function. This energy inlet should not be mistaken for the data inlet of the device. The news and pictures reach the TV through the antenna and the television cable and not through the wall outlet.

Data transfer devices can be linked in a chain just as corresponding devices for energy can be. Many technical data transports have several linked data transfer devices. Fig. 15.6 shows a simplified flow diagram of a live television broadcast in Germany of a football game in Mexico.

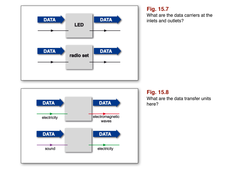

Fig. 15.6
Data transport where the data is transferred five times

According to the examples we have looked at so far, it appears that the final link in every data transfer chain must be a person. This assumption is incorrect though. There are examples of data transports without a person being at the beginning or end. Take, for example, control devices such as a thermostat. In a room of an apartment, there is a temperature sensor, a data source. This tells the boiler whether or not more heating oil should be burned. In this case, the boiler is the data receiver and the data carrier is *electricity.*

If the eyes or ears of a person are at the end of a chain of data transfer links, the term data is commonly replaced by other words such as *news, information, texts, music, pictures, noise, etc.* For many purposes it makes no difference who the addressee is and what the data mean for a person. Therefore, we will always use the word "data."

We have said that "Light, electricity, sound, and radio waves *can be used* as data carriers" but not "Light, *etc. are* data carriers". Whether or not these quantities and substances are to be called data carriers or energy carriers depends upon what one does with them.

The light falling upon a solar collector is called an energy carrier as is the laser light which would be used to drill a small hole. On the other hand, the light going through a fiber-optical guide to broadcast a television show acts as a data carrier. Of course energy is also carried in this case, but it is not important for the application.

Electricity is similar. The electricity flowing in the cable into the house in Fig. 15.1 (or rather, through the house) serves to transport energy. It plays the role of an energy carrier. In contrast, it acts as a data carrier in the telephone line and antenna cable of Fig. 15.2.

Microwaves are another example. In microwave ovens, they serve as energy carriers, and radar uses them as data carriers.

The shock wave generated by blasting a boulder is an example of a sound wave being used as an energy carrier. We are more accustomed to using sound as a data carrier, though.

Even the newspaper which is bought as a data carrier often ends up being an energy carrier when it is burned later.

Exercises

1. Sketch three different data flow diagrams with sources and receivers. (Use different examples than in Fig. 15.4.)

2. Name three different devices that emit data with the carrier sound.

3. Name three devices that receive data with the carrier light.

4. Television sets have a wireless remote control. What is the carrier between this control unit and the television set?

5. In Fig. 15.7, the names of the data carriers at the entrances and exits of the two transfer units are missing. Complete the Figure.

6. In Fig. 15.8, insert the names of the data transfer devices.

7. Sketch a transfer device chain with at least four links. (Use another example than the one in Fig. 15.6.)

Fig. 15.7
What are the data carriers at the inlets and outlets?

Fig. 15.8
What are the data transfer units here?

15.2 The amount of data

A speaks with B, and C speaks with D. A and B talk fast, C and D slowly. More data are exchanged between A and B in the same interval of time than between C and D.

This statement can be made although A and B speak about something entirely different than C and D. This is so because we compare not the meaning of what is being said, but the *amount of data*. In this case, a comparison is simple. However, the problem is often more difficult:

- Does a single page of a fax contain more or less data than a one-minute telephone call?
- Is the amount of data that someone can find within an hour in the Internet greater or smaller than the amount of data in a five-minute telephone call?

In order to answer these questions, we need to be able to measure the quantity of data. In the following, we will get to know a measure for amount of data. We will abbreviate this new quantity with the letter *H*. The unit of amount of data is the bit.

We can use this quantity to specify how many bits are transferred in a one-minute telephone conversation, a one-hour television show, or a smoke signal lasting five minutes.

Later on we will see that this quantity has great practical relevance. Data transfer and data storage cost money and the price for transfer or storage depends upon the amount of data (the number of bits).

We will look at a simple example of a data transfer in order to see what one bit means.

Willy and Lilly plan to take a bicycle trip next Sunday. They need their parent's permission for this. Willy has his permission already but Lilly's parents are only coming home late in the evening on Saturday. Lilly is not allowed out of the house so late in the evening and she should not use the telephone because it would wake up Willy's little sister. How can she let Willy know late on Saturday night if the bicycle trip can take place?

The children have an idea. Their houses are about 300 m apart and the windows of their rooms are within sight of each other. They arrange for Lilly to use a flashlight that can flash red or green at exactly ten o'clock to signal Willy. If she flashes green, she is allowed to go and if she flashes red, she is not.

This kind of data transfer proved itself useful, and Willy and Lilly use it again at other times. The next Sunday evening Willy informs Lilly about the result of a tennis match on TV that she is not allowed to watch.

Willy and Lilly use this method of communication on numerous occasions. At one point Bob, who lives in another house nearby, notices that Willy and Lilly are sending messages to each other. What would Bob be able to say about what is going on? It is impossible for him to know what information is being exchanged, but he knows that each communication is made up of two possible messages because only two signals are being used.

The amount of data exchanged by Willy and Lilly when they communicate about something is always the same: one bit.

The situation can also be described as follows: A question is agreed upon between the data source and the data receiver that can be answered with a yes or a no. For example:

"Are you allowed to go on the bike trip?"

"Did XY win the playoff?"

These kinds of questions are called yes-no questions. In summary:

> 1 bit is the amount of data needed for answering a yes-no question.

Of course, it doesn't matter what kind of sign is used in answering the question. Simply the words "yes" and "no" would suffice. It would be just as easy to replace the red and green lights of the flashlight with blue and white ones, or to make a short and a long flash. It is also possible to transfer the message over a two-core wire using "current on" and "current off".

It is important to realize that the amount of data is independent of the content of a yes-no question. It makes no difference whether the information is important or trivial. The answer carries always 1 bit.

We now return to Willy and Lilly. The amount of data being transmitted between them constantly increases and finally, one evening, Willy has a lot of yes-no questions to answer. They arrange that the light signal for answering the first question should be sent at 22:00 o'clock, with the second one following at 22:05, and the third at 22:10. Because every signal transmits 1 bit, the resulting amount of data is 3 bits. We will now see how several bits can be transmitted with just one signal.

Not every message can be reduced to the answer to a yes-no question. On the contrary, there are more than two answers possible for most questions. Willy and Lilly found this out.

On two consecutive days, a captivating two part criminal drama was shown on TV. After the first episode, it was clear that one of the following four people had committed a murder:

the cleaning lady
the mailman
the murder victim's sister
the husband of the victim.

Willy and Lilly are anxious to know who will be revealed as the murderer. Now Lilly finds out that the next evening her aunt is coming to visit and the TV will be turned off. Willy must somehow let her know who the murderer is.

There is a problem to solve though. There are four possible answers, and not just two, for the question "Who is the murderer?" The problem is quickly solved. Willy proposes the following procedure: He will get a blue plastic film and using this with his flashlight, he can generate four different colors. These are red, green, blue and white. They could agree upon the following assignment:

cleaning lady:	green
mailman:	red
sister:	white
husband:	blue

This kind of agreement is called a code. Willy's idea will certainly work, but Lilly has another idea. "Two colors are enough for us. You can use the first one to tell me if it is a man or a woman and the second one to let me know if the murderer is a relative of the victim or not." Lilly's code looks like this:

cleaning lady:	green – green
mailman:	red – green
sister:	green – red
husband:	red – red

This kind of code, where only two signs or characters are used, is called a *binary code*.

Willy and Lilly have been discovering an important theorem of data technology. In the following, we will formulate it more generally. For now we will graphically represent Lilly's method and draw a *decision tree*, Fig. 15.9.

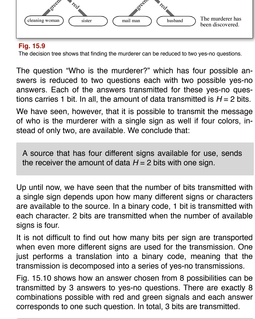

Fig. 15.9
The decision tree shows that finding the murderer can be reduced to two yes-no questions.

The question "Who is the murderer?" which has four possible answers is reduced to two questions each with two possible yes-no answers. Each of the answers transmitted for these yes-no questions carries 1 bit. In all, the amount of data transmitted is $H = 2$ bits.

We have seen, however, that it is possible to transmit the message of who is the murderer with a single sign as well if four colors, instead of only two, are available. We conclude that:

> A source that has four different signs available for use, sends the receiver the amount of data $H = 2$ bits with one sign.

Up until now, we have seen that the number of bits transmitted with a single sign depends upon how many different signs or characters are available to the source. In a binary code, 1 bit is transmitted with each character. 2 bits are transmitted when the number of available signs is four.

It is not difficult to find out how many bits per sign are transported when even more different signs are used for the transmission. One just performs a translation into a binary code, meaning that the transmission is decomposed into a series of yes-no transmissions.

Fig. 15.10 shows how an answer chosen from 8 possibilities can be transmitted by 3 answers to yes-no questions. There are exactly 8 combinations possible with red and green signals and each answer corresponds to one such question. In total, 3 bits are transmitted.

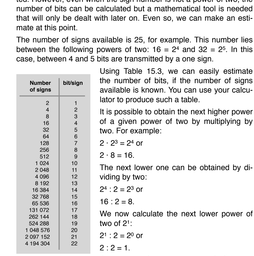

Fig. 15.10
Decision tree for three consecutive yes-no questions

If 8 different signs are available, the 3 bits can be transmitted with just one sign.

We see how this rule can be generalized: If a source has 16 different signs at its disposal, 4 bits can be transmitted with one sign. If the number of available signs is 32, then there are 5 bits per sign, etc.

> If the number of signs is 2^n, the number of bits transmitted per sign is n.

When the number of available signs is equal to a power of the *number* two, we can easily specify how many bits per sign are transmitted. However, even when the sign number is not a power of two, the number of bits can be calculated but a mathematical tool is needed that will only be dealt with later on. Even so, we can make an estimate at this point.

The number of signs available is 25, for example. This number lies between the following powers of two: $16 = 2^4$ and $32 = 2^5$. In this case, between 4 and 5 bits are transmitted by a one sign.

Number of signs	bit/sign
2	1
4	2
8	3
16	4
32	5
64	6
128	7
256	8
512	9
1 024	10
2 048	11
4 096	12
8 192	13
16 384	14
32 768	15
65 536	16
131 072	17
262 144	18
524 288	19
1 048 576	20
2 097 152	21
4 194 304	22

Table 15.3

Using Table 15.3, we can easily estimate the number of bits, if the number of signs available is known. You can use your calculator to produce such a table.

It is possible to obtain the next higher power of a given power of two by multiplying by two. For example:

$2 \cdot 2^3 = 2^4$ or

$2 \cdot 8 = 16$.

The next lower one can be obtained by dividing by two:

$2^4 : 2 = 2^3$ or

$16 : 2 = 8$.

We now calculate the next lower power of two of 2^1:

$2^1 : 2 = 2^0$ or

$2 : 2 = 1$.

It is possible to raise a number to the zero power. Any number raised to the power of zero yields 1 (the only exception is zero itself, the expression 0^0 is undefined). In particular, $2^0 = 1$. Now we can complete Table 15.3. If the number of available signs is 1, each sign transmits 0 bits. Is this surprising? Actually, it isn't. When Willy and Lilly arrange for Willy to send an agreed upon sign at 22:00 o'clock, Lilly will not find out anything new when she receives it.

How does this work in the following case? Willy and Lilly arrange that he will send a white flash signal at ten o'clock in the evening if their favorite team has won a certain game. If they lose the game, he won't send any signal. A message is definitely being sent even when apparently only one sign is being used. Actually, two signs are in use here. The flashlight can be either turned on or off at ten o'clock so the signs are either 'light' or 'dark'.

The situation is very similar with school bells, doorbells, automobile horns, sirens, the blinking lights at a train crossing, etc. In all of these cases, there are two signs. For example "the school bell is ringing" and "the school bell is not ringing" or "the light is blinking" and "the light is not blinking".

Data transfer and data encoding are two rather similar processes; to distinguish between them is somewhat arbitrary. We will look at Willy and Lilly's two codes again. It is entirely possible to encode messages coming in with Willy's code of four colors of light into Lilly's code of two colors. This process can be graphically represented exactly like a transfer process, Fig. 15.11.

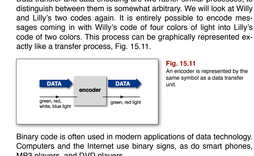

Fig. 15.11
An encoder is represented by the same symbol as a data transfer unit.

Binary code is often used in modern applications of data technology. Computers and the Internet use binary signs, as do smart phones, MP3 players, and DVD players.

15.3 Examples of a data transport

Morse code

Figure 15.12 shows how telegrams used to be transmitted. Source and receiver are part of an electric circuit; they are connected by two wires. The source is just a switch that can close the circuit. When

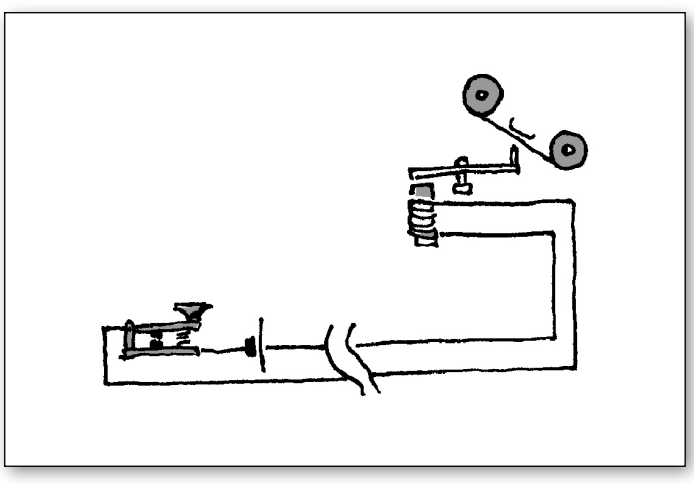

Fig. 15.12
Transmission of Morse code

the circuit is closed, an electromagnet at the receiving end pushes a stylus against a strip of paper moving by. The Morse code was used for this kind of data transfer, Fig. 15.13. This code used four signs: “dot” (short closing of the circuit), “dash” (longer closing of the circuit), “short pause” (within a letter) and “long pause” (between two letters). Morse code has four signs, so 2 bits per sign are transmitted. The Morse code is still used in navigation and amateur radio.

·—	a	·—··	l	·——	w	—··—·	/
—···	b	——	m	—··—	x	·————	1
—·—·	c	—·	n	—·——	y	··———	2
—··	d	———	o	——··	z	···——	3
·	e	·——·	p	·—·—·—	period	····—	4
··—·	f	——·—	q	——··——	comma	·····	5
——·	g	·—·	r	··——··	?	—····	6
····	h	···	s	———···	:	——···	7
··	i	—	t	—·—·—·	;	———··	8
·———	j	··—	u	·—·—·	+	————·	9
—·—	k	···—	v	—····—	-	—————	0

Fig. 15.13
Morse code

Writing

One of the most important methods for storing and transporting data is writing. How many bits are in a written character? First we must determine how many different characters there are. These include upper and lower case letters, numbers, punctuation marks, arithmetic operators, and other special characters. The space between two letters represents a character. We will assume that we only use characters that are found on the keyboard of a computer. The keyboard has about 45 keys. Each key has a dual function. This means that, depending upon whether or not the shift key has been pressed, one of two different characters can be written. The letter keys have the possibility of upper or lower case letters. In all, the keyboard can write 90 characters, or between 2^6 and 2^7 characters. Each character carries just about 7 bits.

Pictures

A computer creates pictures on its screen. Which amount of data must the computer send to its screen to do this? On a typical computer screen, a picture is made up of approximately $1600 \cdot 1200 = 1\ 920\ 000$ points, or so called pixels. If we have a black and white screen, each pixel is either black or white. One bit for each pixel is sent from the computer to the screen. This comes to 2 Mbit total, for all the pixels. Computers can also produce color pictures whereby a pixel can receive 16 777 216 different colors. Because 16 777 216 is equal to 2^{24}, the computer must transmit 24 bits for each pixel. For the entire picture, this means

$24 \cdot 2\ \text{Mbit} \approx 50\ \text{Mbit}$.

Before a picture is “saved” the data is usually “compressed”. This is possible because in a typical picture there are whole regions that have the same color. When stored in the so-called jpeg format a picture can be compressed to between a tenth and a hundredth of its original amount of data.

The quantity of data of a measured value

When someone measures something, he or she gets data about the object upon which the measurement is performed.

Consider a balance scale that can be used up to 5 kg. In the set of weights, the smallest one is a 1 g weight. Therefore, the scale can answer the question “How heavy is the object?” in 5000 different ways. The number of available signs is 5000, and the amount of data associated with the answer is around 12 bits. A modern analytical balance supplies up to 20 bits per weighing.

In order to calculate the amount of data in an *analog*-scale, for example, a thermometer, the first thing to clarify is the resolution, i.e., the difference between neighboring values on the scale that can still be distinguished. The thermometer has an accuracy of reading of about 1 °C. If the measurement range goes from –30 up to 90 °C, the resulting number of available signs is 120. When the temperature is read, 7 bits are received.

Exercises

1. There are approximately 100 000 different postal area codes (in the area served by the German postal service). What is the amount of data carried by one area code?

2. The amount of data carried by a telephone number depends upon whether it is dialed locally, nationally, or internationally. Estimate the amount of data in a telephone number of a local telephone network with 10,000 connections.

3. The Chinese language has numerous different characters. Normally, only about 2000 of these are used. How many bits does one character have, based upon this number?

4. A source transmits 5 bits with each sign. How many different signs does the source have at its disposal?

5. A source has 3 different signs available. Draw a decision tree for this source. (It should contain three consecutive decisions.) Estimate the amount of data that reaches the receiver with three consecutive signs from this source.

6. Source A has a number of available signs that agrees with a power of two. Source B has twice as many signs. What is the consequence of this for the amount of data being transmitted per sign by each source?

7. A magic card trick:

Any 16 cards of a normal playing deck can be used. The magician lets someone from the audience pick one. The person looks at the card so the magician cannot see it. The magician puts it back with the others and they are all shuffled. Now the magician turns them over one by one. In doing so, he creates four piles. One card on the first pile, the second card on the second pile, etc. He repeats the process until all 16 cards are lying on the table. Now the audience member must say which pile his card is in. The magician then puts all four piles into one and repeats the process of putting them into four piles. Again, the person from the audience must say which pile his card is in. Now the magician knows which card it is. He again puts the four piles together and turns them over one by one until the chosen one appears.

Which amount of data does the magician need to receive in order to identify one out of sixteen cards? How many bits does he receive each time the audience member indicates which pile his card is in? How does his trick work?

15.4 Data currents

Whenever something flows, whether it is water, cars, people, energy, electricity or any other things, substances or physical quantities, we can ask about the strength of the current. As we know, the strength of a current can be obtained by measuring the amount of the quantity that flows in a given time past a given place and dividing it by the time this quantity needs to flow by.

The term *data current strength* or simply *data current* at a given point on its transmission line is understood to be the ratio of the amount of data flowing by that point and the time needed to flow by.

$$\text{data current} = \frac{\text{quantity of data}}{\text{time span}}$$

$$I_H = \frac{H}{t}$$

The unit for data currents I_H is bit/s. The more bits that are transmitted per second, the stronger the data current is.

A data current is a useful quantity. It allows for comparison of performance between data transmission devices. We will look at some examples of data currents.

Telephone and radio

When one makes a telephone call, a data current of about 50 kbit/s flows. The quality of acoustic data by telephone is not very high. A better transmission requires a larger data current. This is why the data current received from a CD is much higher than that of a telephone. It is around 1000 kbit/s. Also acoustic data can be reduced by smartly coding them and omitting information which we would not perceive anyway. MP3 coding reduces the amount of data to about one tenth.

Television

In section 15.3, we calculated the amount of data for one still picture on a television screen. The result was H = 50 Mbit. A show on TV transmits about 25 pictures per second. This is how we get the impression of continuously moving objects on the screen. The strength of the data current flowing from the television station to the receiver results from these values. It is:

50 Mbit/picture · 25 pictures per second = 1250 Mbit/s
≈ 1000 Mbit/second.

Again this data flow can be reduced by an appropriate coding.

In any case we have:

> The data current that flows for visual perception, is about 1000 times as great as that for acoustical perception.

The Internet

When one gets information from the Internet, the average data current is shown on the computer screen. It depends upon how heavily used the network is at that moment and where the server transmitting the data is located. As we know, this can vary greatly.

Exercises

1. A letter or symbol written on the keyboard carries about 7 bits. What is the data current flowing from the keyboard to the computer when 180 keystrokes per minute are being typed?

2. You are downloading information from the Internet. The data current is 2400 bit/s. How long does it take to receive the text of a DIN-A4 page? (Assume there are 40 lines per page, 70 characters per line, and 7 bits per character.)

16

Electricity and Electric Currents

In the same way that mechanics deals with momentum and its transfer, and thermodynamics deals with heat and heat transfer, the science of electricity is about electricity and its currents.

What is electricity? For the moment, we can give only a very general answer. The more you work through this and the following chapters, the better you will understand electricity. For now, we will answer the question like this: Electricity is what “flows” in the wires of a cable of an electric device when it is turned on. Electricity can be understood as a kind of substance that can get from place to place similarly to the “substances” of momentum and entropy that we have already dealt with.

We can usually see whether or not a body has momentum because we can see whether or not it is moving. We can also “see” if a body has a little or a lot of entropy because we can notice if it is cold or hot. However, we have no sensory organs with which to determine the electricity of a body. An electric shock lets us feel electricity, but we want to avoid this because it is dangerous.

You already know that electricity plays an important role in technology. In the chapters coming up, you will get to know how some technical devices work.

The technical applications of electricity can be divided into two main categories.

One kind of application has to do with using electricity for the transfer and storage of energy. Electricity is a practical energy carrier. Many electrical devices serve to transfer energy to another energy carrier or to transfer energy from another carrier to electricity. Examples of these appliances are electric motors, generators, and all electric heaters.

A second type of technical application for electricity is the transfer, storage and processing of data. Music, written and spoken texts, pictures, numbers and other symbols are examples of data. This field of technology is called electronics.

At first glance, electricity doesn’t appear to play much of a role in nature with the exception of lightning. However, this appearance is deceiving. In fact, the structure of the micro-world of atoms and molecules is governed to a great extent by electricity. The structure of atoms is the result of electricity, and their cohesion is made possible by electricity. This is what atomic physics deals with.

16.1 The electric circuit

Fig. 16.1 shows a light bulb connected via a switch to a battery. A flashlight is constructed like this. The energy goes from the battery to the lamp with the help of the carrier electricity. It is transferred there to the carrier light. The energy comes out of the battery, reaches the lamp, and then leaves it with the light. In the process, the battery slowly drains, meaning that its energy content diminishes.

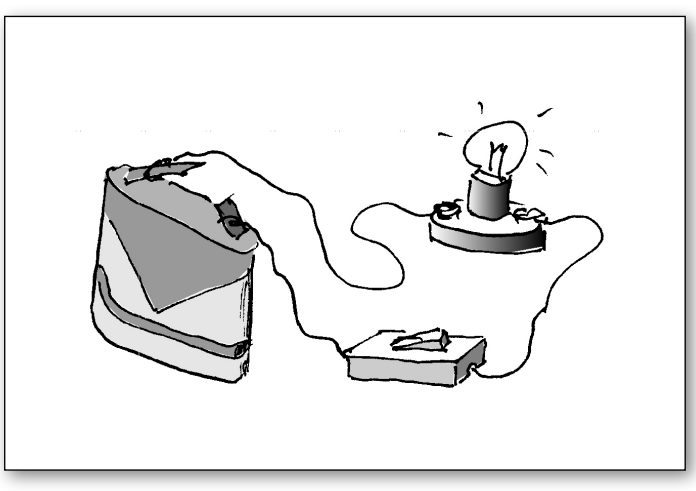

Fig. 16.1
Circuit of a flashlight

The carrier of the energy, i.e., the electricity, takes another path. It flows in a circuit. It comes out of one of the two battery terminals, the positive terminal, and flows through the wire to the light bulb and through its filaments. It flows on through the second wire over the switch to the battery's negative terminal, and then through the battery to the positive terminal again. Because the electricity moves in a closed path without building up anywhere, this configuration is called an *electric circuit*. The current of electricity is called the *electric current* for short.

Electricity cannot flow through just any material. The materials that allow it to flow easily are called electric *conductors.* Materials through which electricity cannot flow are called *insulators.* All metals are conductors. Insulators include air, glass, and most plastics.

Electricity is a physical quantity. The symbol for this quantity is *Q*. It is measured in Coulombs, abbreviated to C.

An electric circuit is very similar to a hydraulic circuit used, for example, in a power shovel, Fig. 16.2.

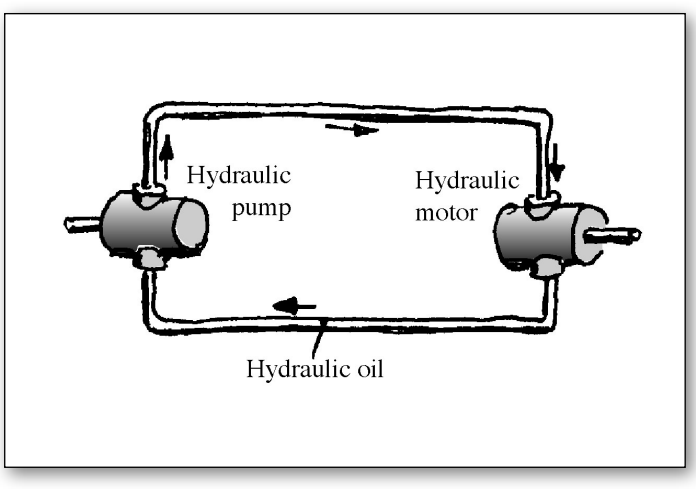

Fig. 16.2
A hydraulic circuit has great similarity to an electric circuit.

Here as well, the energy carrier, i.e., hydraulic oil, is flowing in a closed circuit. The flow diagrams in Fig. 16.3 and 16.4, clearly demonstrate this similarity.

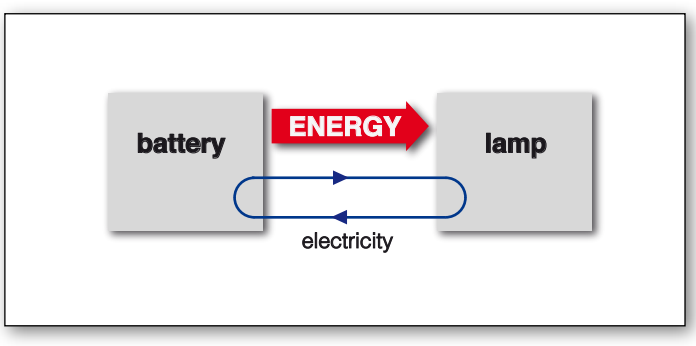

Fig. 16.3
Flow diagram of the electric circuit of Fig. 16.1

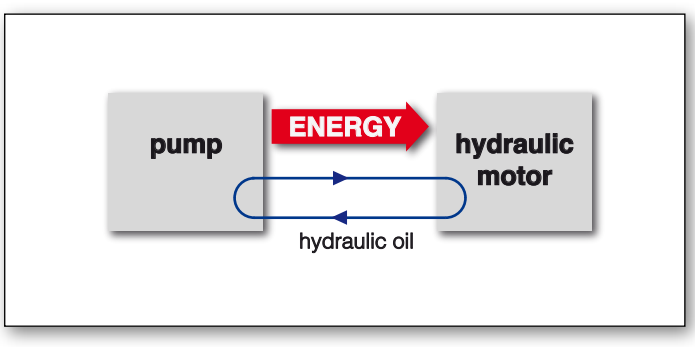

Fig. 16.4
Flow diagram of the hydraulic circuit of Fig. 16.2

Similarly to how the pump in a hydraulic circuit makes the liquid flow, the battery makes the electricity flow in our electric circuit. We can consider the battery as a kind of *electricity pump*.

There are other sources that supply energy with the carrier electricity, i.e., other kinds of electricity pumps. One of these is the bicycle dynamo. The same device is also found in cars. Very large dynamos such as the ones in electric power plants are called generators. Solar cells are still another type of electricity pump. While a dynamo receives its energy with angular momentum, solar cells receive their energy with light.

Batteries, dynamos, and solar cells are electricity pumps.

The circuit in Fig. 16.1 is first interrupted or "open". We close the switch and electricity flows through the lamp. Where does this electricity come from? From the battery, you might think, exactly like the energy. In fact, this is not so. Similarly to a water pump that can only release as much water as it takes in, an electricity pump can only emit as much electricity at its positive terminal as it receives at its negative terminal. So, where does the electricity come from?

It is in the components of the electric circuit right at the outset: in the battery, in the lamp, and in the wires. This electricity is there naturally; it is not put there by anyone. Every piece of wire, every piece of metal contains electricity that begins to flow when the wire or piece of metal is connected to a circuit.

When one builds an electric circuit, one needs not worry about filling it with electricity. It is as if one would have a hydraulic circuit where the hydraulic oil were already in the pumps, hoses and motors. Such circuits can begin to work immediately. They don't need to be filled with oil first.

In the following sections, we will deal frequently with electric circuits—sometimes very complicated ones. Therefore, it makes sense to introduce symbols for their representation. Fig. 16.5 shows the symbols of a battery, an open switch, a light bulb and an electric motor. A wire (an electric conductor) is represented by a simple straight line.

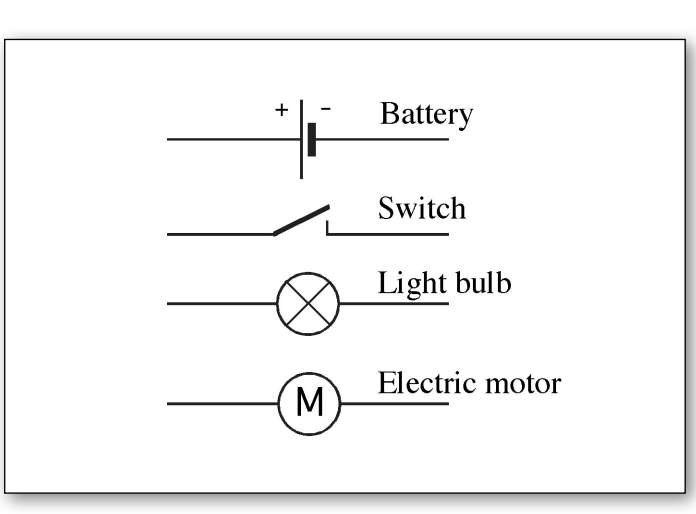

Fig. 16.5
Symbols of some electric components: battery, open switch, light bulb, and electric motor

In Fig. 16.6, the circuit of Fig. 16.1 is represented with the help of these symbols, once with an open switch and once with a closed switch.

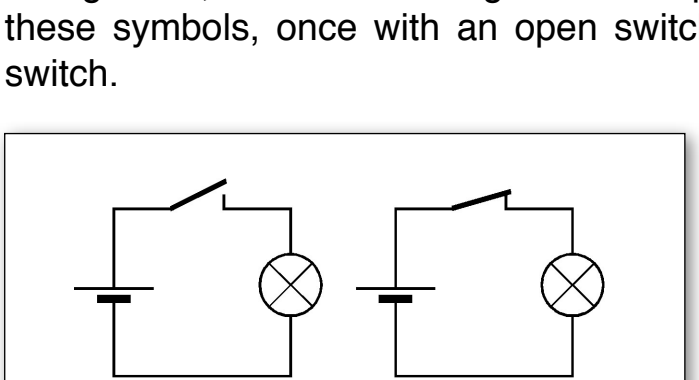

Fig. 16.6
Symbolic representation of the circuit of Fig. 16.1 with the switch at two different positions

16.2 Electric current

We consider a point P of an electric circuit, Fig. 16.7. A lot or a little electricity can flow by this point per second, depending upon what kind of battery and what kind of lamp we are using. We say that *the strength of the electric current*, or simply the *electric current*, can be greater or smaller. Similarly to other currents such as energy currents or water currents, we define

$$\text{electric current} = \frac{\text{electricity}}{\text{time span}}$$

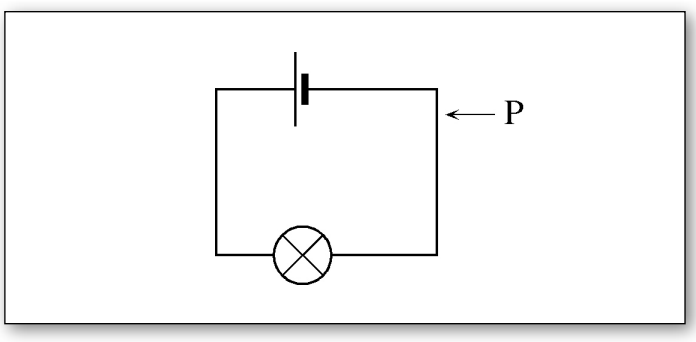

Fig. 16.7
A certain amount of electricity per second flows past location P of the circuit.

The electric current is abbreviated by the letter *I*. Therefore

$$I = \frac{Q}{t}$$

The unit of electric currents is

Coulomb/second = C/s.

A simpler name is normally given to this derived unit. This name is the Ampere, abbreviated to A. Therefore

Ampere = Coulomb/second, or A = C/s.

Table 16.1 recapitulates the names of the new quantities together with their units and the corresponding abbreviations.

Table 16.1

Name of quantity	Amount of electricity	Electric current
Abbreviation	*Q*	*I*
Name of unit	Coulomb	Ampere
Abbreviation	C	A

In order for us to understand which currents are strong and which ones are weak, we will take some measurements. The device used for measuring electric currents is called an *ammeter*. An ammeter has two terminals, Fig. 16.8.

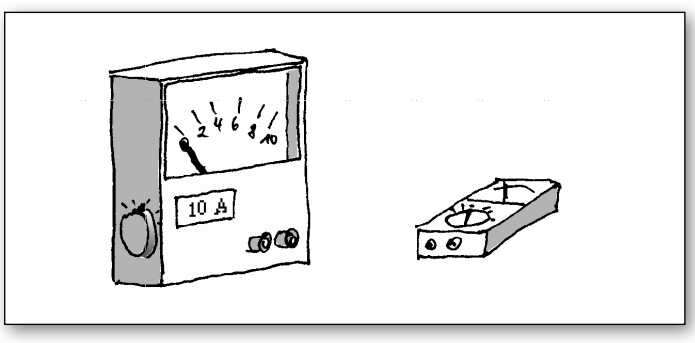

Fig. 16.8
Two ammeters

In order to measure the strength of the current in the wire in Fig. 16.9a, the wire is cut through, Fig. 16.9b. Two new ends are created in this way. These ends are connected to the ammeter's two terminals, Fig. 16.9c. The electricity must now flow through the ammeter.

In order to measure the electric current in a wire, the wire must be cut through, and the two new ends must be attached to the terminals of the ammeter.

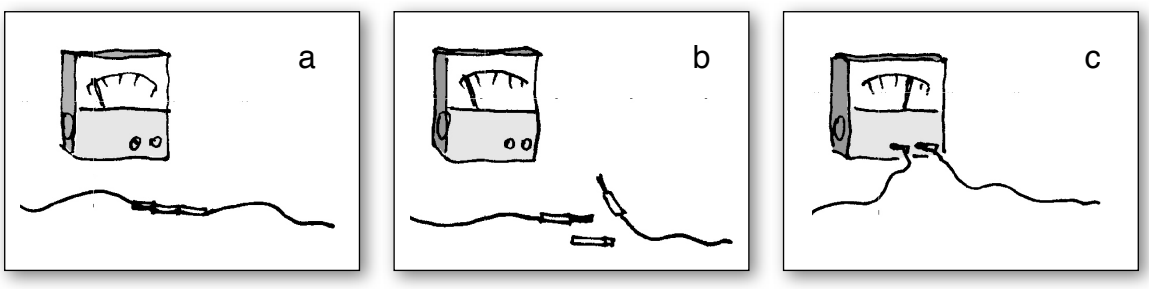

Fig. 16.9
Measuring the electric current in a wire

Fig. 16.10 shows a circuit with an ammeter built into it (the symbol for the ammeter is a circle with the letter A). The ammeter shows 0.5 A, a typical value for a small lamp.

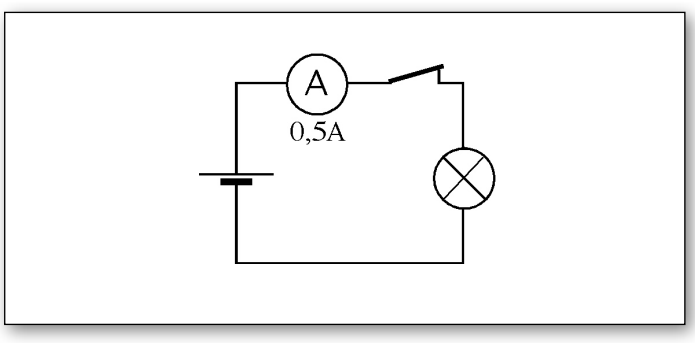

Fig. 16.10
The electric circuit of Fig. 16.1 with an ammeter built in

If the ammeter is introduced at another location of the circuit, it will, naturally, show the same value, Fig. 16.11. For every second, the same amount of electricity must flow at every point of the circuit, i.e., through every cross section of the wire.

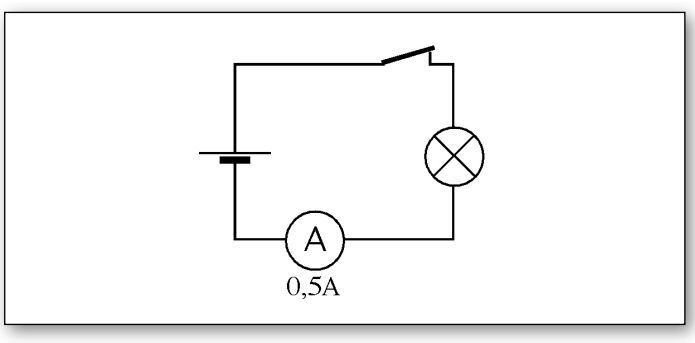

Fig. 16.11
The ammeter always shows the same value no matter where it is in the circuit.

It is possible to attach several ammeters in a circuit without changing anything about the currents, Fig. 16.12. Each of these Ammeters shows 0.5 A. This is similar to simultaneously measuring with three stop watches the time someone would need to run one hundred meters.

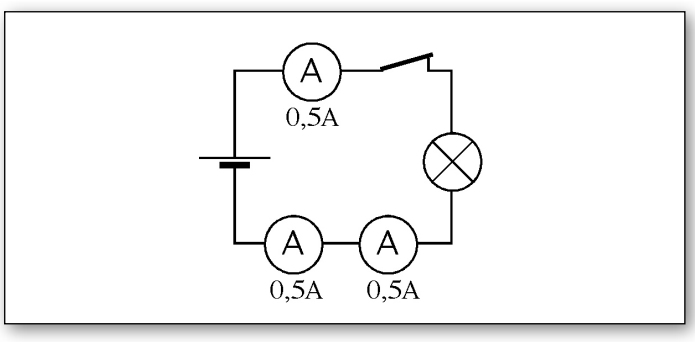

Fig. 16.12
Several ammeters "in series" show the same as a single one.

Table 16.2 shows some typical values of electric currents.

Table 16.2

Electric currents	
through a 75 W light bulb	0,34 A
in the conductors of a pocket calculator	0,01 mA
through the motor of an electric train engine	500 A
in a bolt of lightening	several 1000 A
through a toy motor	1 A

16.3 The junction rule

The point where several currents meet is called a *junction*. It does not matter what kind of currents they are. They could be energy currents, water currents, or electric currents.

Fig. 16.13 shows an arrangement of electric components creating a more complex circuit. This kind of arrangement is called a branching circuit.

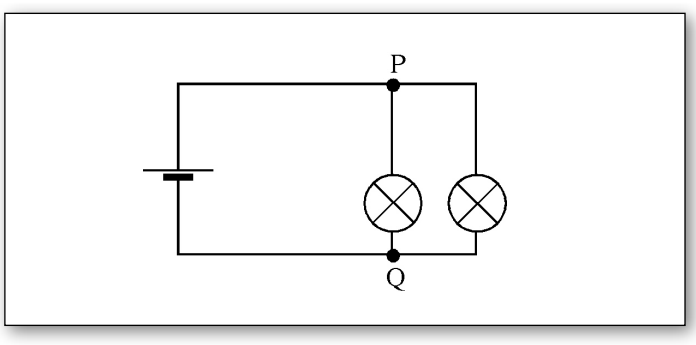

Fig. 16.13
The branching circuit contains the junctions (nodes) P and Q.

The circuit in Fig. 16.13 contains two junctions: junction P and junction Q. The electricity comes from the positive terminal of the battery. The electric current branches at junction P. One part of it flows through the lamp on the left, and the rest flows through the one on the right. At junction Q, the two currents join again. All the electricity flows from Q to the battery's negative terminal, through the battery, and back to the positive terminal.

Fig. 16.14 shows the same setup as Fig. 16.13, but here three ammeters have been introduced. The ammeter that measures the electric current I_1 in conductor 1, i.e., before the branching point P, reads 4 Ampere. This means that 4 Coulomb per second are flowing towards node P. The ammeter in conductor 2 shows 2.5 A. This means that 2.5 C per second is flowing away from node P through cable 2. What does the third ammeter show? For a proper balance, 1.5 C per second must flow away from P through conductor 3. The current in conductor 3 is 1.5 A.

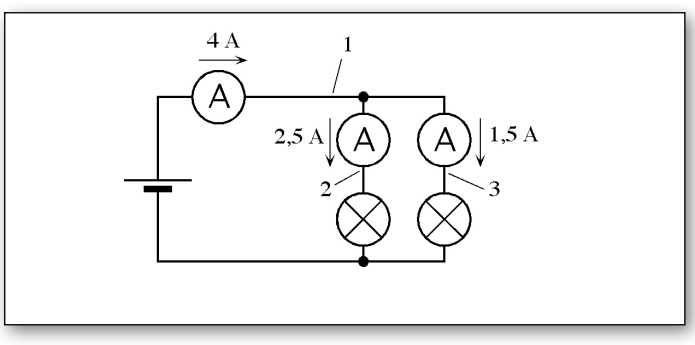

Fig. 16.14
Three ammeters are built into the circuit of Fig. 16.13.

It is the same situation when two rivers join, Fig. 16.15. In this case, as well, the same amount must flow away from the junction as flows into it.

Fig. 16.15
The junction rule also holds for merging rivers.

Fig. 16.16 shows a part of a more complicated electric circuit. In this case, 6 conductors meet at a junction. Check to see whether the balance is correct.

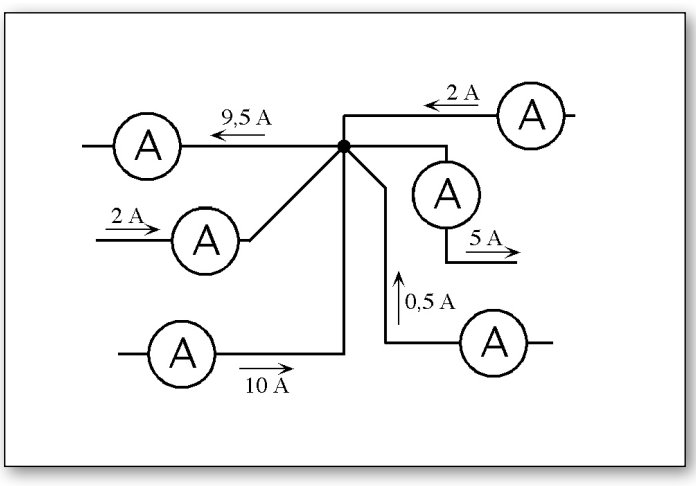

Fig. 16.16
A section of a complicated circuit. Six cables meet at one junction.

We have been using our well-known junction rule the entire time:

> The currents flowing into a junction are, in total, equal to the ones flowing away from it.

Exercises

1. What is the electric current flowing at point P in Fig. 16.17a? What direction is it flowing in?

2. What is the electric current flowing at point P in Fig. 16.17b? What direction is it flowing in?

3. What can be said about the currents at points P and Q in Fig. 16.18a?

4. (a) Introduce two switches into the circuit in Fig. 16.18b so that the lamps can be switched on and off independently. (b) Introduce a single switch into the circuit by which both lamps can be turned on and off together.

5. What do the ammeters 2, 3, and 4 read in Fig. 16.19a?

6. What is the electric current at P in Fig. 16.19b? Fit an ammeter into the circuit that can measure the current through the motor, and another one that measures the current through the lamp.

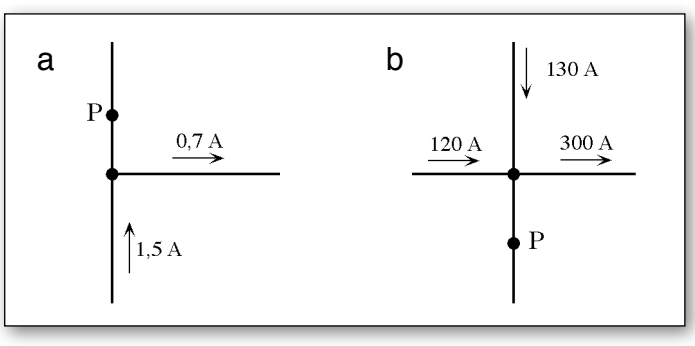

Fig. 16.17
For Exercises **1** and **2**

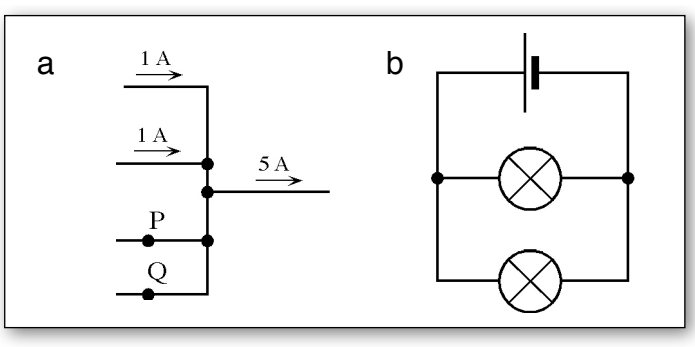

Fig. 16.18
For Exercises **3** and **4**

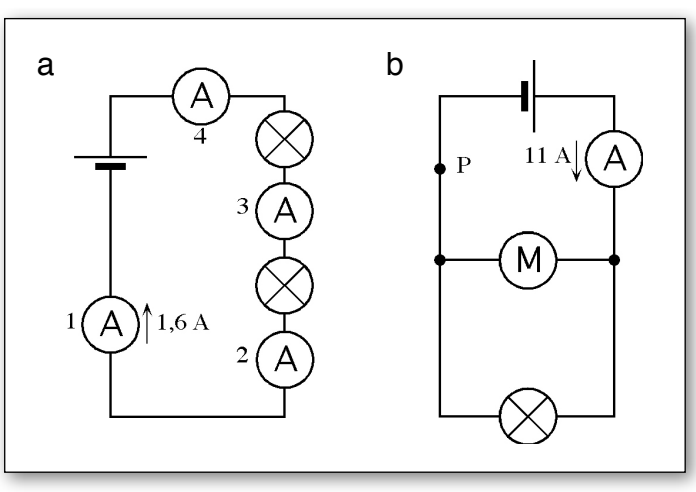

Fig. 16.19
For Exercises **5** and **6**

16.4 The electric potential

A water pump causes the water at its exit to have a higher pressure than at its entrance, Fig. 16.20. It creates a pressure difference. This

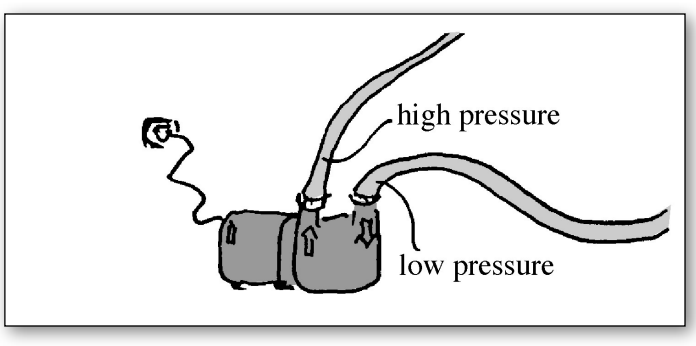

Fig. 16.20
The Pressure at the outlet of the water pump is higher than at its inlet.

pressure difference can drive a water current.

A battery, i.e., an electricity pump, sets up a driving force for an electric current. We introduce a quantity which has a higher value at the positive terminal than at the negative terminal, Fig. 16.21. This

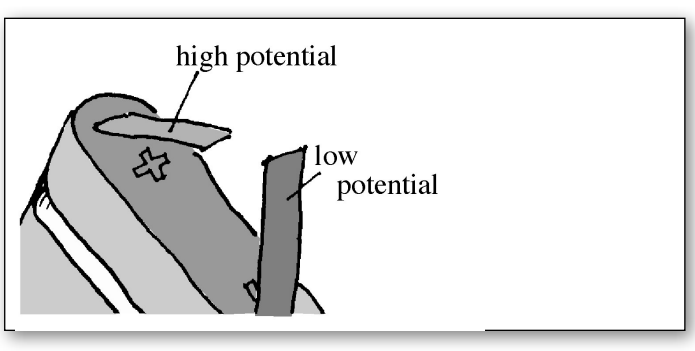

Fig. 16.21
The electric potential of the battery's plus terminal (outlet) is higher than at its minus terminal (inlet).

quantity is called *electric potential*. The electric potential of an electric circuit corresponds to the pressure in a hydraulic circuit.

An electricity pump (battery, dynamo) creates a potential difference. This potential difference is a driving force for an electric current.

The potential is higher at the plus terminal than at the minus terminal.

A battery produces a potential *difference*, and this potential difference is the driving force for an electric current.

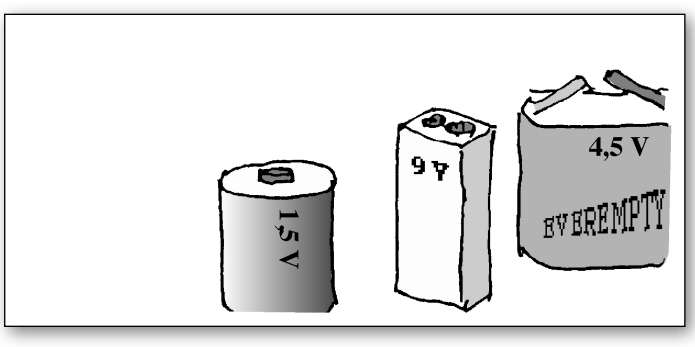

Fig. 16.22
Three "electricity pumps" with their potential difference values printed on them.

Fig. 16.22 shows some electricity pumps: three different types of batteries. The potential difference of each of these energy sources is usually printed on them.

The unit for measuring the potential is the Volt, abbreviated to V. A mono-cell creates a potential difference of 1.5 V, a battery produces a potential difference of 4.5 V, and a car battery 12 V.

Potential difference is called voltage.

The word *voltage* is often used instead of potential difference. We can say that we have a voltage of 4.5 V between the terminals of a flat battery.

The Greek letter ϕ (pronounced phi) is used as the symbol for potential and U is used for potential difference or voltage. For our battery we have

$\phi_+ - \phi_- = 4.5$ V,

or

$U = 4.5$ V.

Names, units, and abbreviations of these quantities are contained in Table 16.3.

Table 16.3

Name of quantity	Electric potential	Voltage
Abbreviation	ϕ	U
Name of unit	Volt	Volt
Abbreviation	V	V

It is not necessary to look at what is printed on a battery in order to know its potential difference because voltage is easy to measure by using a voltmeter. A voltmeter has two terminals (like an ammeter). In order to measure the voltage between two points of a circuit, both points are connected to the terminals of the voltmeter, Figs. 16.23 and 16.24.

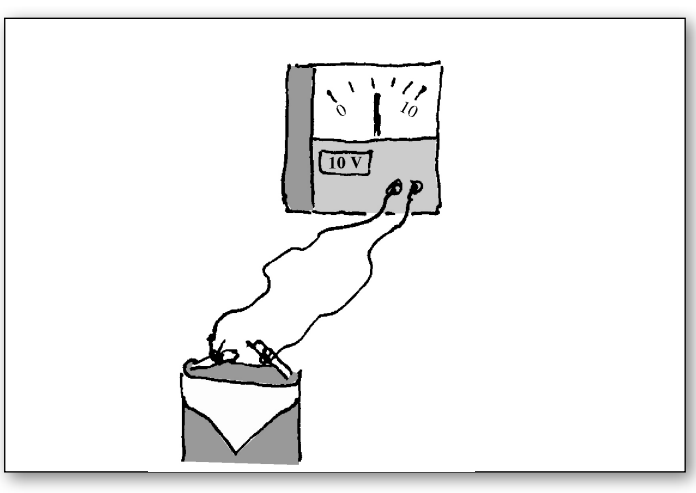

Fig. 16.23
In order to measure the voltage between two different points, they are connected to the terminals of a voltmeter.

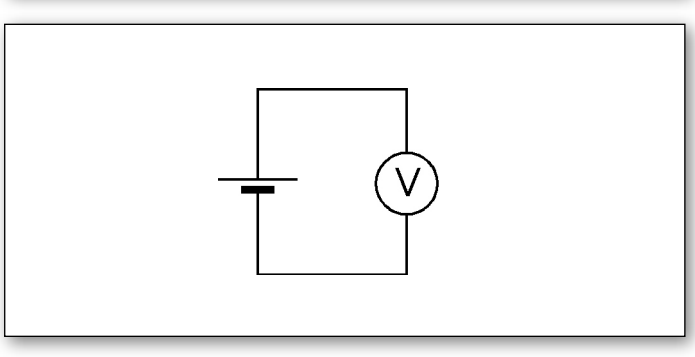

Fig. 16.24
The same arrangement as in Fig. 16.23, represented by symbols

Locations that are connected to each other by a cable, have the same potential. This is why the four voltmeters in Fig. 16.25 all show the same voltage.

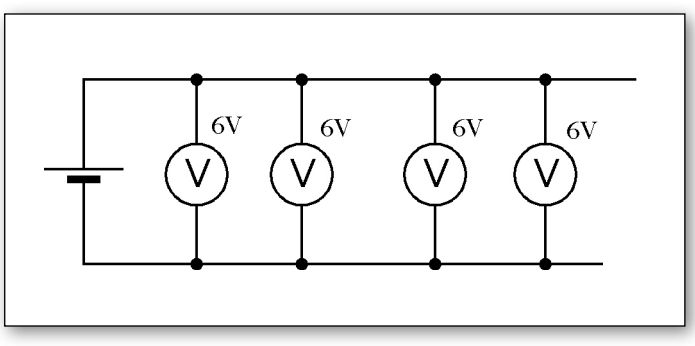

Fig. 16.25
Several voltmeters "in parallel" show the same voltage as just a single one would.

Voltmeters are built so that only a very small electric current can flow through them. An ammeter connected to the wire of a voltmeter therefore shows 0 A, unless it is very sensitive, Fig. 16.26.

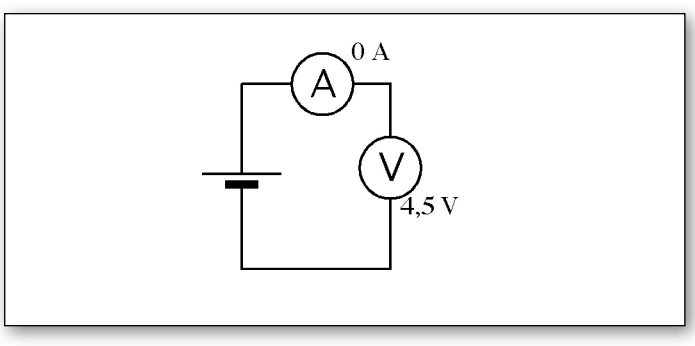

Fig. 16.26
(Almost) no electric current flows through the voltmeter, so the ammeter shows 0 A.

When a battery is empty (when all the energy is gone), it cannot produce any potential difference. A voltmeter can be used to find out whether a battery is still usable.

A *power supply* is an electric energy source that must be plugged into an outlet. The power supply gets its energy with the carrier electricity, and also emits this energy with electricity. There is, however, a difference between the inlet and the outlet: The voltage at the outlet is not the same as at the inlet. The voltage at the outlet is often adjustable. In addition, the voltage between the terminals of the socket, i. e. the inlet of the power supply is alternating. An alternating voltage's values change very quickly with time. At their outlet most power supplies have direct voltage which stays constant with time.

16.5 The zero point of the electric potential

You have a battery on a table in front of you. The potential difference between its terminals is 4.5 V. The potential at the plus terminal is therefore 4.5 V higher than at the minus terminal. What is the potential of the minus terminal? What is the potential of the plus terminal?

These questions are not simple to answer. It might be easier to solve this problem if we answer another question first. Fig. 16.27 shows a one-meter ruler standing vertically upon a table. We ask the question: At what height is the upper end of the ruler? All we really can say is that the upper end is one meter above the lower end. But how high is the lower end? The answer depends upon what we use as the reference for heights. Do we use the floor of the room? The level of the ground outside the house? Or maybe another level? You probably already know that the height of terrain is based upon sea level. Sea level has been arbitrarily set at 0 m. We could indicate the height of the upper end of the meter ruler based upon sea level, but such heights are generally difficult to determine.

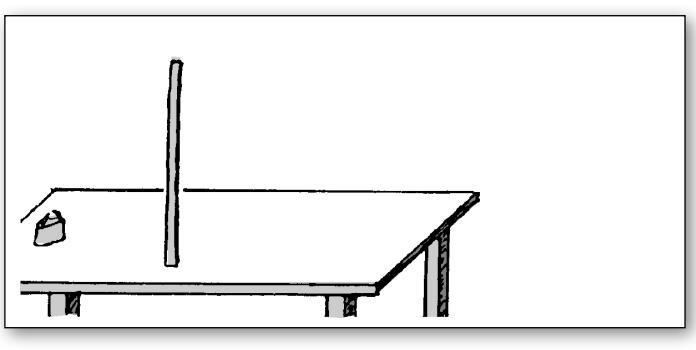

Fig. 16.27
What is the potential of the positive terminal of the battery? At what height is the upper end of the meter stick actually?

Electric potential behaves similarly to height. First we must determine which electric conductor we will give the potential 0 V. Using this as the basis, all the other wires, terminals, etc. can be given values of the electric potential. The conductor we use as the reference potential should be available to everyone. A conductor that fulfills this requirement is the Earth, so the following has been decided:

> The Earth's potential is 0 V.

If any point of an electric circuit is connected by a wire to the Earth, we can be sure that this point will be at 0 Volts. The point has been *grounded* or earthed.

In order to ground something, it is not even necessary to connect a conductor to the Earth, because you can also get zero potential from the wall outlet. Indeed, the contact in the round hole below the two vertical slots of the outlet is grounded, Fig. 16.28.

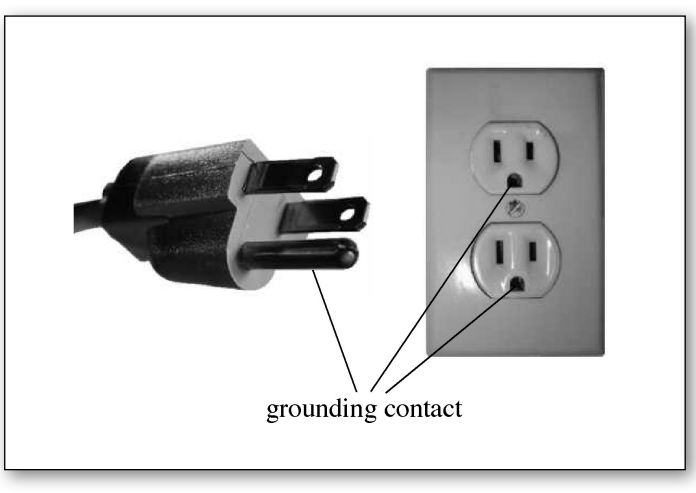

Fig. 16.28
The grounding contact of the electric outlet is in the hole below the two slots.

Now, back to the battery on the table. Based upon what we have just said, we do not know the individual values of the plus and minus terminals, just as we do not know exactly how high the ruler on the table was. We can, however, easily clarify the situation for the battery by just grounding one of the terminals. Fig. 16.29a shows a battery whose minus terminal is grounded (notice the symbol for the Earth). We have

$\phi_- = 0$ V.

Therefore, the plus terminal is at

$\phi_+ = 4.5$ V.

The plus terminal in Fig. 16.29b is grounded, so in this case

$\phi_+ = 0$ V

and

$\phi_- = -4.5$ V.

The potential of the minus terminal is now negative. In both cases (Figs 16.29a and 16.29b), we naturally have

$\phi_+ - \phi_- = 4.5$ V.

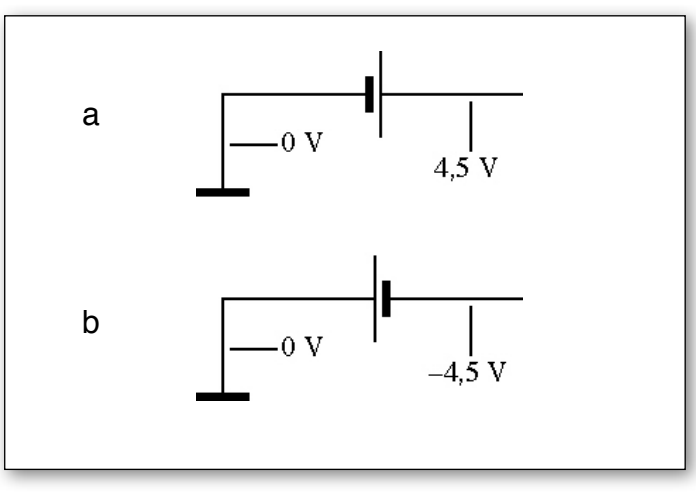

Fig. 16.29
(a) The battery's minus terminal is grounded, the plus terminal is at a potential of +4.5 V.
(b) The plus terminal of the battery is grounded. The minus terminal has a potential of –4.5 V.

The expressions "plus terminal" and "minus terminal" have become everyday expressions, but they are somewhat misleading. They lead us to believe that a plus terminal is always at a positive potential and a minus terminal is always at a negative potential. This is not necessarily the case, as seen in Fig. 16.29. In Fig. 16.29a, the minus terminal has the potential 0 V. Its potential is not negative. In Fig. 16.29b, the plus terminal does not have positive potential. This can be seen more clearly in Fig. 16.30. In the Figure, a 9 V battery and a 1000 V power supply are connected in series. The plus terminal of the power supply is grounded, so its potential is 0 V. Its minus terminal lies 1000 V lower, meaning at –1000 V. Because the plus terminal of the battery is connected to this point, this battery terminal is also at a potential of –1000 V. This means that the potential of the plus terminal of the battery is negative.

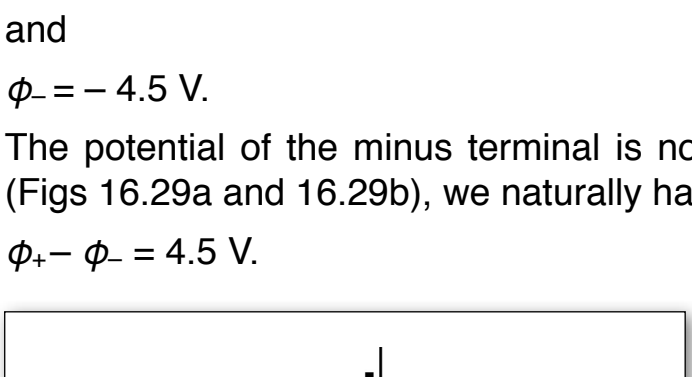

Fig. 16.30
The battery's positive terminal is at a potential of minus one thousand volts.

Fig. 16.31 shows a circuit which is grounded at one point.

Fig. 16.31
A circuit where one point is grounded

Exercises

1. The batteries in Fig. 16.32a each create a voltage of 4.5 V. What are the potentials at points 1, 2 and 3?

2. Each of the batteries in Fig. 16.32b produces a potential difference of 12 V. What are the potentials at points 1, 2 and 3?

3. Two 9 V batteries are connected in series, Fig. 16.33a. What readings do the three voltmeters show?

4. Sketch a voltmeter into Fig. 16.33b that shows the voltage between the terminals of the lamp. Draw a voltmeter that measures the battery's voltage.

5. Give some examples of circuits that cannot be grounded.

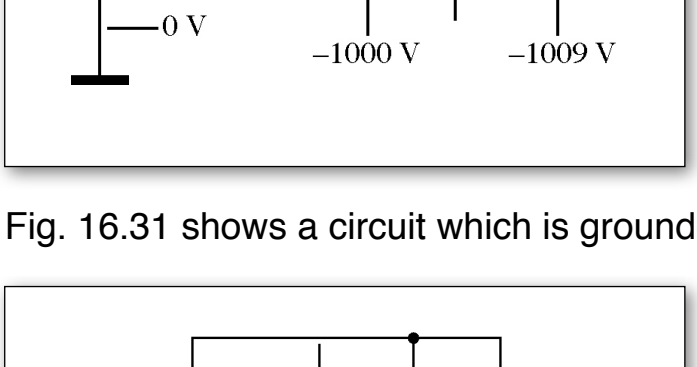

Fig. 16.32
For Exercises **1** and **2**

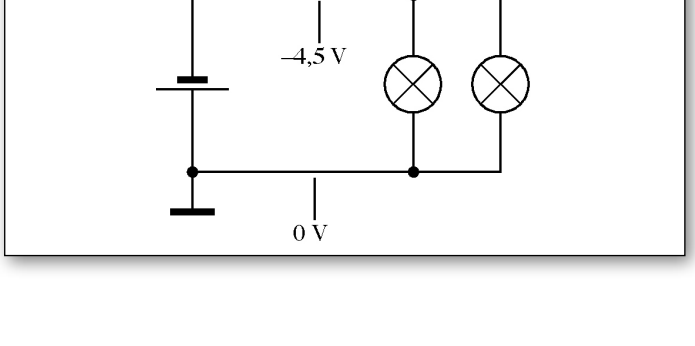

Fig. 16.33
For Exercises **3** and **4**

16.6 Driving force and currents

We connect an electric motor with a 6 V battery and then with a 9 V battery, Fig. 16.34. In the second case, the motor runs more quickly than in the first. The ammeter shows that the electric current is stronger in the second case than in the first.

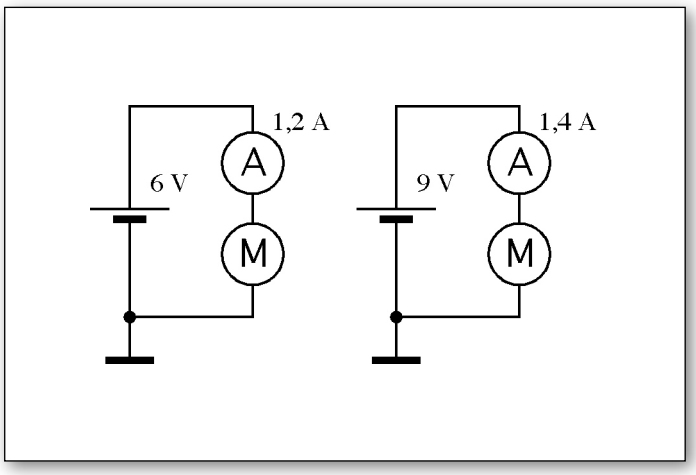

Fig. 16.34
The higher the voltage across the motor, the stronger the current flowing through it.

We connect a lamp to an adjustable power supply and slowly increase the voltage. The higher the voltage, the stronger the current flowing through the lamp.

The experiments demonstrate what you have probably expected: The higher the voltage, the stronger the current.

> The greater the electric potential difference between two points (the greater the driving force), the stronger the electric current flowing from one point to the other.

We have two differently constructed little bulbs. The same battery is connected first to one, and then to the other, Fig. 16.35. We find that the current flowing through one lamp is stronger than the one flowing through the other lamp. Apparently, the lamps do not have the same *resistance* to the current.

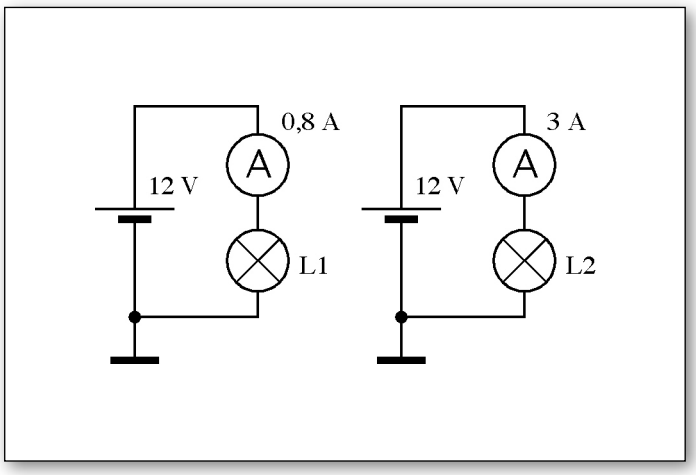

Fig. 16.35
The lamp in the circuit on the right has a smaller resistance than the one on the left.

We see that the current flowing through a device (through a lamp or a motor, for example), depends upon two things:

– the voltage between the terminals of the device;
– the resistance of the device to the current.

> The strength of an electric current flowing through a device is greater,
> – the greater the potential difference between the terminals of the device;
> – the smaller the resistance of the device to the current.

16.7 Applications

We will now learn a method that makes solving problems of technical applications of electricity easier.

Whenever the circuit diagram of an electrical setup is drawn, the conductors are sketched in color. All the conductors with the same potential have the same color. It is obvious that one continuous wire will have one color. When it goes through an electric device (lamp, motor, battery, dynamo, etc.), the color will usually change.

Figs. 16.36 to 16.38 show some examples of this.

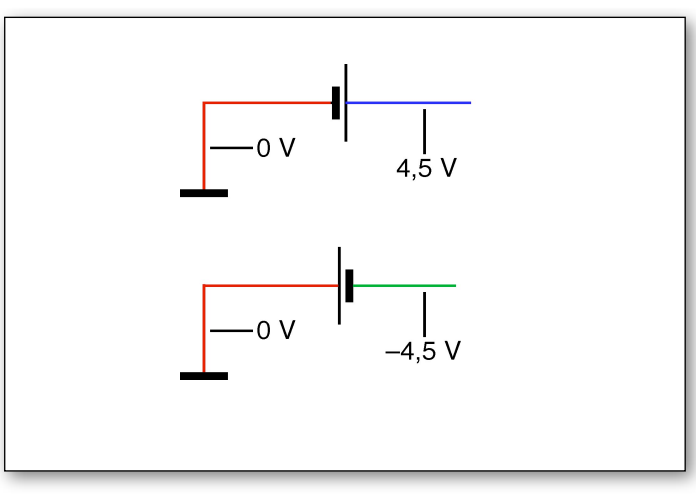

Fig. 16.36
Differently colored lines in the conductor stand for different potentials.

Fig. 16.36 shows the battery of Fig. 16.29 with its cables drawn in this new way. Fig. 16.37 shows, once again, the lamps of Fig. 16.31 connected in parallel. Fig. 16.38 shows a circuit with four different potential values.

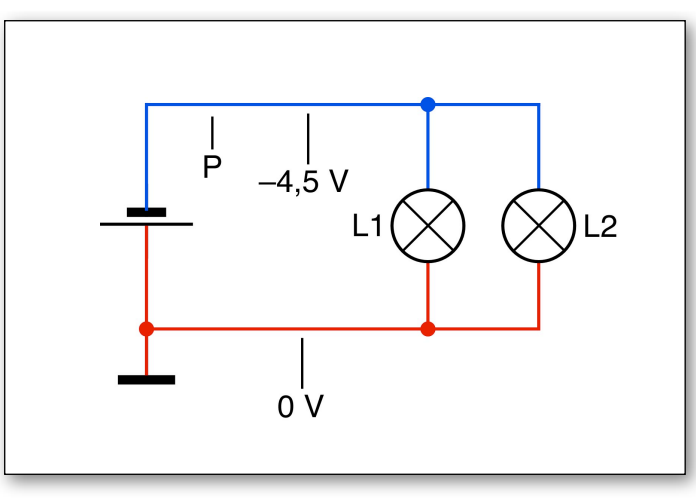

Fig. 16.37
The circuit of Fig. 16.31. The potentials are represented by differently colored lines.

We want to apply the color labeling of the conductors to two problems.

1. Lamps 1 and 2 in Fig. 16.37 are built identically. A current of 3 A is flowing at point P. What is the current through lamp 1 and what is it through lamp 2?

The junction rule is valid here for the branching points, so

$I_{L1} + I_{L2} = 3$ A.

(I_{L1} und I_{L2} are the currents through the lamps.) We now see by the colors of the conductors that the voltage across both lamps is the same (the same one as across the battery). The driving force of the electric current in both lamps is the same. Because both lamps are constructed identically, the currents in both of them must be identical, so that,

$I_{L1} = I_{L2} = 1{,}5$ A.

2. Section B of the conductor in Fig. 16.38 is at a potential of 6 V. The lamps 1, 2 and 3 are identical. What is the voltage produced by the battery?

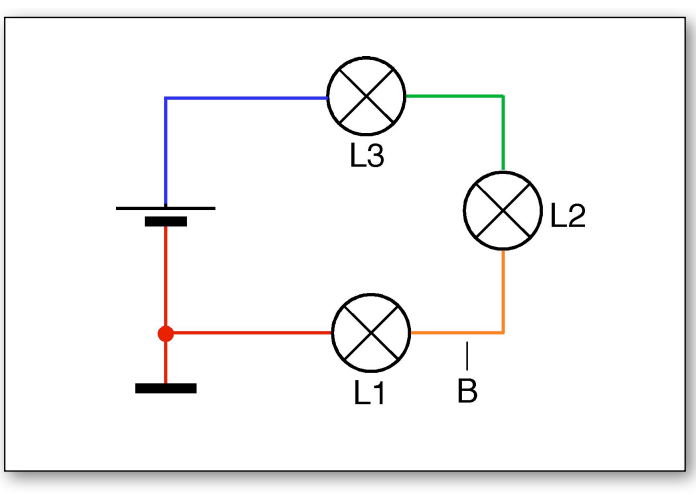

Fig. 16.38
There are four different values of the potential in this circuit.

The current is the same everywhere because there is no branching of the circuit. The voltage across lamp 1 is 6 V. It represents the driving force for the current through lamp 1. The same current also flows through lamps 2 and 3 (which are identical to lamp 1). Therefore, the driving force needed for the electricity to flow through them is the same one needed for lamp 1, and this is 6 V. If one moves from the battery's plus terminal, through the three lamps, and to the minus battery terminal, it goes in three 6 V steps down to 0 V. The plus terminal must, in this case, be at 18 V.

In both of these examples, the potential at the inlet of a lamp was different from the one at the outlet. However, this is a rule that does not always apply. A lamp with no electric current flowing through it has the same potential at the inlet and the outlet, otherwise a current would be flowing. Fig. 16.39 shows two examples.

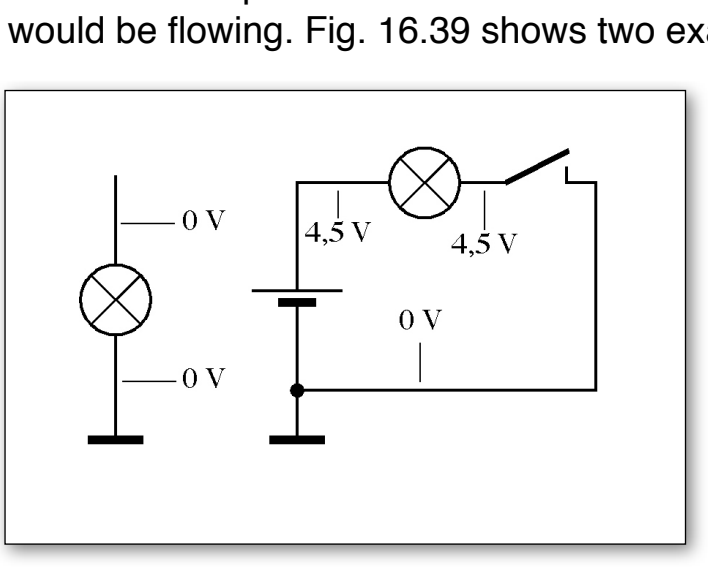

Fig. 16.39
The lamps have no current flowing through them, so their terminals must have equal potentials.

Exercises

1. The batteries in Fig. 16.40a are 4.5 V batteries. Color-label the conductors according to their potentials and give the values of the potential for all the conductor sections.

2. The current flowing through the battery in 16.40b is 1.6 A. Color-label the conductors. What is the current through the lamps?

3. The electric potential at point C in Fig. 16.41 is 20 V. The three lamps are identical. Color-label the conductors. Give the potentials of the conductor sections A, B, and D. What voltage does the battery set up? What happens to the potentials when the switch is opened?

4. The battery voltage in Figs. 16.42a and 16.42b is 12V. The lamps are identical.

Color-label the conductors.

What is the value of the potential at point P?

What are the potential differences across lamps L1 and L2?

Is the current through L1 stronger when the switch is closed (Fig. 16.42a) or when it is open (Fig. 16.42b)?

Is the current stronger in lamp L2, when the switch is open or closed?

5. The voltage of the power supply in Figs. 16.43a and 16.43b is 150 V. The lamps are identical.

Color-label the conductors.

Give the potentials of all the sections of the conductor.

Which lamp burns even when the switch is opened?

6. The batteries in Figs. 16.44a and 16.44b set up a voltage of 9V. The lamps are identical. Color-label the sections of the conductors and give the potentials for them.

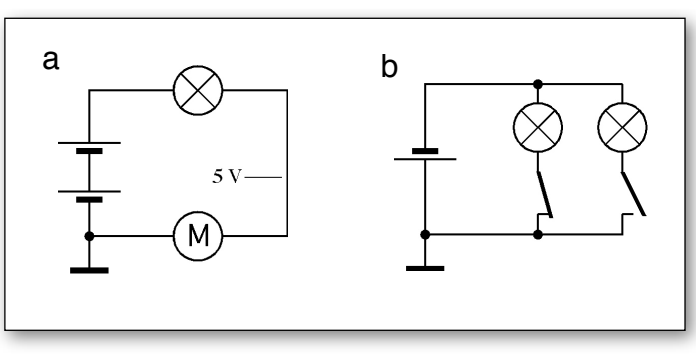

Fig. 16.40
For Exercises **1** and **2**

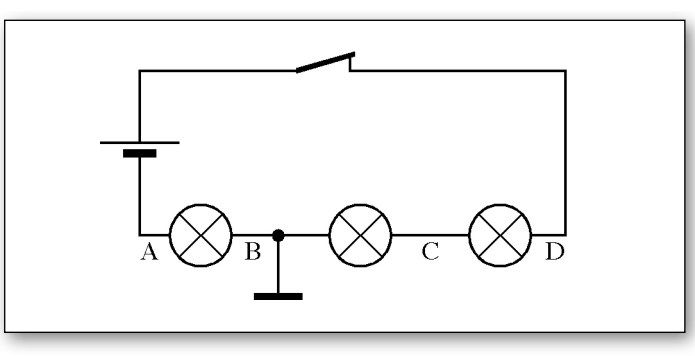

Fig. 16.41
For Exercises **3**

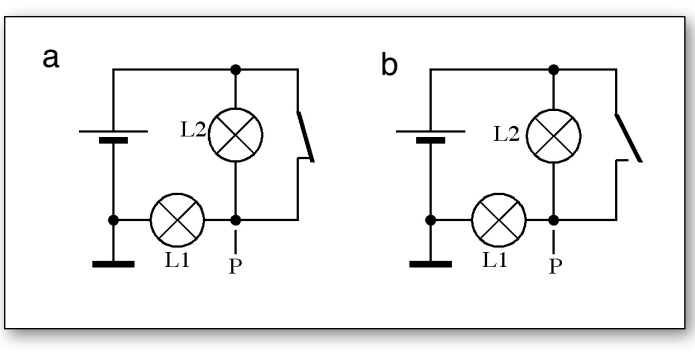

Fig. 16.42
For Exercises **4**

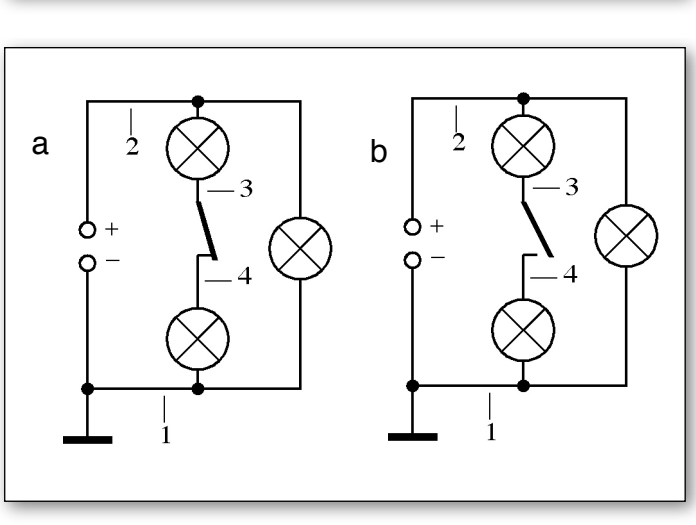

Fig. 16.43
For Exercises **5**

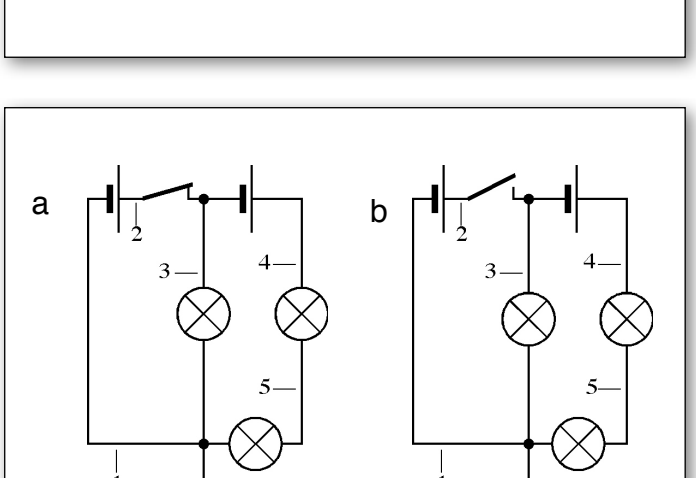

Fig. 16.44
For Exercises **6**

16.8 Electric resistance

When we want to have an electric current flow through an object, we set up a voltage across it to create a driving force. Every object, however, tends to hinder such a current. It resists the electricity flowing through it. One says that it *has* a resistance.

Some objects have a great resistance, meaning that they conduct an electric current badly or not at all. Others have hardly any resistance; they are good electricity conductors.

Electric cables, for example, have only a slight resistance. This does not mean that they have no resistance at all.

How the electric current flowing through an object reacts to the voltage applied can be a complicated matter. When the voltage is increased, the strength of the current usually increases, but this is not always the case.

We wish to investigate the relation between voltage and current for various electrical devices. Fig. 16.45 shows how to do this: We plug the object we wish to investigate into a power supply with adjustable voltage. This voltage can be read on the setting knob. (If you don't trust the readings on this knob, it is possible to measure the voltage independently.) The electric current caused by the voltage is shown on an ammeter. We give the voltage different values and read the current for each of these values.

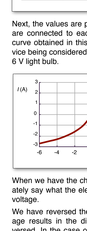

Fig. 16.45
Measurement of characteristic curves: Various values of voltage are set up and in each case, the current caused by the voltage is measured.

The pairs of numbers obtained in this way are first put into a table. Fig. 16.46 shows such a table of values for a 6 V light bulb.

U/V	I/A	U/V	I/A	U/V	I/A
−6	−2,7	−2	−1,6	3	2
−5	−2,5	−1	−1	4	2,3
−4	−2,36	0	0	5	2,5
−3	−2	1	1	6	2,7
		2	1,6		

Fig. 16.46
Table of values for the characteristic curve of a small light bulb

Next, the values are put into a *U-I* coordinate system and the points are connected to each other by the smoothest possible line. The curve obtained in this manner is the *characteristic curve* of the device being considered. Fig. 16.47 shows the characteristic curve of a 6 V light bulb.

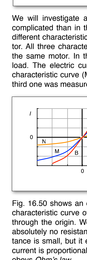

Fig. 16.47
Characteristic curve of a light bulb. The corresponding table in shown in Fig. 16.46.

When we have the characteristic curve of a device, we can immediately say what the electric current is that flows through it at a given voltage.

We have reversed the voltage in our light bulb. Reversing the voltage results in the direction of the electric current also being reversed. In the case of the light bulb, there is a point symmetry between the positive and the negative parts of the curve.

The characteristic curve of a *diode* is shown in Fig. 16.48. We see that the curve has no point symmetry. In case you do not know what a diode is used for, you might think about it with the help of the characteristic curve. The curve shows that the diode allows the electric current to flow only in one direction. It has the same function for an electric current that a bicycle tire valve has for air flow.

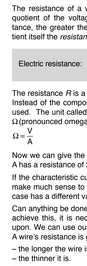

Fig. 16.48
The characteristic curve of a diode is not point symmetrical.

We will investigate an electric motor. The situation here is more complicated than in the previous cases. This is because there is a different characteristic curve depending upon the load upon the motor. All three characteristic curves in Fig. 16.49 were generated by the same motor. In the first one (N), the motor ran freely, with no load. The electric current constantly remained small. The second characteristic curve (M) shows a medium value of the load, and the third one was measured with the axle blocked (B).

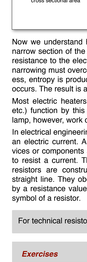

Fig. 16.49
The characteristic curves of an electric motor depend upon the load attached to the motor. (N: no load, M: medium load, B: axle blocked)

Fig. 16.50 shows an especially simple characteristic curve. It is the characteristic curve of a long wire. It has the form of a straight line through the origin. We have assumed up until now that a wire has absolutely no resistance. You see now that this is not so. The resistance is small, but it exists. The characteristic curve shows that the current is proportional to the voltage applied. One says that the wire obeys *Ohm's law.*

Ohm's law: $I \sim U$

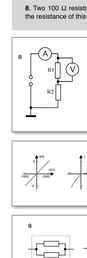

Fig. 16.50
Characteristic curve of a long wire. The wire follows Ohm's law.

The characteristic curves of two different wires A and B is shown in Fig. 16.51. When the same driving force is applied, the current in wire B is weaker than that in wire A. B has larger resistance than A.

Fig. 16.51
Characteristic curves of two wires with different resistances

The resistance of a wire can be characterized by establishing the quotient of the voltage and current. The greater the wire's resistance, the greater the quotient is. For this reason, we call the quotient itself the *resistance* of the wire. We label it with the letter *R*.

Electric resistance: $R = \frac{U}{I}$

The resistance *R* is a physical quantity. Its unit is Volt/Ampere (V/A). Instead of the composite expression Volt/Ampere, the word *Ohm* is used. The unit called Ohm is further abbreviated to the Greek letter Ω (pronounced omega). Therefore

$$\Omega = \frac{V}{A}$$

Now we can give the resistance of the two wires of Fig. 16.51. Wire A has a resistance of 2 Ω and wire B has a resistance of 5 Ω.

If the characteristic curve of a device is not a straight line, it doesn't make much sense to calculate a quotient of *U/I*. The quotient in this case has a different value for each point on the characteristic curve.

Can anything be done to reduce the resistance of a wire? In order to achieve this, it is necessary to know what the resistance depends upon. We can use our experience with water hoses to help us here. A wire's resistance is greater

– the longer the wire is and
– the thinner it is.

The resistance also depends upon what material the wire is made of. When wires of the same length and thickness are compared, we see that wires made of silver and copper have the lowest resistance. They conduct electric currents equally well. An aluminum wire, however, has about twice the resistance, and an iron wire has about six times the resistance of a copper wire.

In Fig. 16.52, the relation between electric current, potential difference, and the properties of the conductor are represented schematically.

Fig. 16.52
Relationship between electric current, *potential difference* and properties of the conductor.

Now we understand how a light bulb works. It is essentially just a narrow section of the circuit, a section of the conductor with a great resistance to the electric current. The electricity flowing through this narrowing must overcome a kind of frictional resistance. In the process, entropy is produced just as in any other process where friction occurs. The result is a rise in the temperature of the wire.

Most electric heaters (hot plates, irons, the heater in a hairdryer, etc.) function by this principle. A microwave oven or a fluorescent lamp, however, work differently.

In electrical engineering or electronics it is often desirable to obstruct an electric current. A resistance is required. For this purpose, devices or components are produced that have no other function than to resist a current. These components are called *resistors*. These resistors are constructed so that their characteristic curve is a straight line. They obey Ohm's law, and they can be characterized by a resistance value or Ohm number. Fig. 16.45 shows the circuit symbol of a resistor.

For technical resistors $I \sim U$ is valid.

Exercises

1. A voltage of 20 V is applied to an unknown resistor. A current of 4 mA is measured. What is the resistance of the resistor?

2. A voltage of 120 V is applied to a 2 kΩ resistor. What is the electric current flowing through the resistor?

3. An electric current of 0.1 mA flows through a 1 MΩ resistor. What is the voltage across the resistor?

4. The power supply in Fig. 16.53a produces a voltage of 35 Volts. The Ammeter shows 5 A, and the voltmeter 10 V. What is the resistance *R1*? What is the voltage across resistor *R2*? What is the resistance *R2*?

5. The voltage of the battery in Fig. 16.53b is 12 V. Each resistor has a resistance of 100 Ω. Give the potentials of each conducting section. What are the voltages across the three resistors? What are the electric currents flowing through the three resistors? What is the electric current flowing through the battery?

6. You find some old electronic components in a box. You have no idea what they were used for. You take the characteristic curves for three of these components and find the relationships represented in Fig. 16.54. What are these components? Be as exact as you can.

7. Two 100 Ω resistors are connected in parallel, Fig. 16.55a. What is the resistance of the entire set up? Formulate a rule.

8. Two 100 Ω resistors are connected in series, Fig. 16.55b. What is the resistance of this entire set up? Formulate a rule.

Fig. 16.53
(a) For Exercise **4**, (b) For Exercise **5**

Fig. 16.54
For Exercise **6**

Fig. 16.55
(a) For Exercise **7**, (b) For Exercise **8**

16.9 Short circuits and fuses

Fig. 16.56a shows two terminals of a battery that are connected directly to each other by a wire. In Fig. 16.56b, the two conductors of a cable leading to a motor touch each other. In both cases, electricity flows directly from one terminal to the other terminal of the power supply without taking a detour over an electrical device that would act as an energy receiver. This situation is called a *short circuit*.

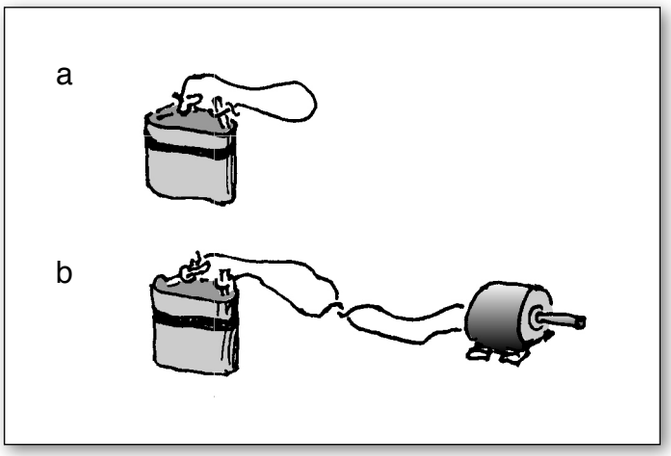

Fig. 16.56
Short circuit: The electricity does not take the 'detour' through thae energy receiver (load).

The circuit produced by the short has a very small resistance. Therefore a very strong current flows. The electric current can be so strong in a short circuit that it becomes dangerous. The conductors can become overheated and start smoldering.

The strength of the current depends upon its power source. A flat battery, a monocell, or similar sources are not dangerous because the currents there are only a few Amperes. An automobile battery is a different story, though. The current of a short circuit can, in this case, be as much as several hundred Amperes. The short circuit current of a wall outlet would be even higher if the building's main fuses didn't prevent a very strong current.

The function of a fuse is to interrupt the electric current as soon as it exceeds a given value. In homes, the maximum value is usually 15 A.

When a battery short circuits, it empties. In this case, "empty" means empty of energy. Where does this energy go? As we already know, the conductors get warm. Energy is needed to warm up the wires (to produce entropy in the wires). This takes only a part of the energy lost by the battery. Moreover, if the battery was shorted by a wire with very little resistance, the wire will not really heat up much. So where did the energy go?

If you are willing to sacrifice a monocell, you can try this yourself. Connect the terminals with a short thick wire. The wire will not get warm – but the battery will. During a short circuit, entropy is produced in the energy source itself. The energy leaves the source, not with the carrier electricity, but with the carrier entropy.

16.10 Alternating current

We construct a somewhat unusual electric energy source from two batteries and a toggle switch. We connect it to a lamp, Fig. 16.57. We flip the switch at regular time intervals of 3 seconds, for example. In doing so, the lower conductor stays at a constant 0 V. The upper one jumps back and forth between + 4.5 V and – 4.5 V.

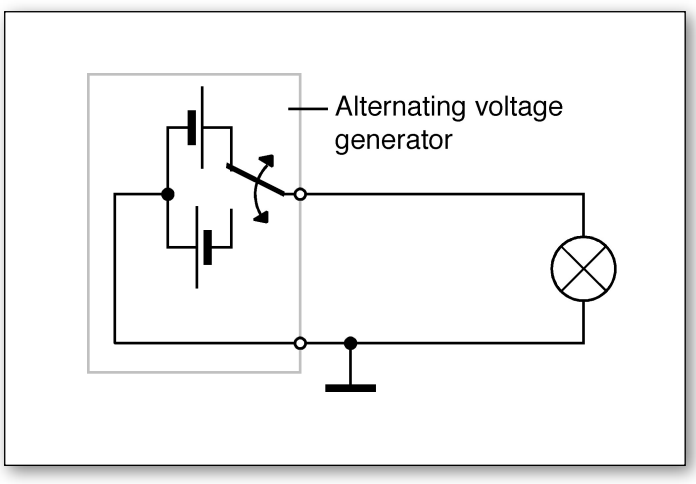

Fig. 16.57
Alternating voltage generator. One terminal always is grounded. The potential of the other terminal alternates between positive and negative values.

Fig. 16.58 shows the potential of the upper conductor as a function of time. There is a so-called *alternating voltage* across the light bulb. This alternating voltage drives the electric current back and forth through the light bulb. There is an *alternating current* flowing through it.

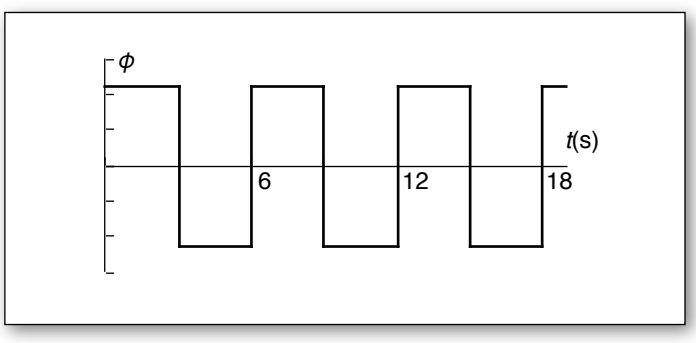

Fig. 16.58
The potential of the upper conductor in Fig. 16.57 as a function of time

There is an alternating voltage between the terminals in the slots of a typical outlet. One of these terminals is always grounded. (It is at the same potential as the grounding contact below the slots.) The potential of the other terminal alternates between positive and negative. There are some differences to the home-made source of Fig. 16.57, though:

– The potential of the outlet's non-grounded terminal alternates much more quickly, at 120 times per second. It is positive 60 times per second and negative 60 times per second. The so-called *frequency* is 60 Hertz.
– The potential of the non-grounded terminal does not change suddenly, but steadily as in waves, as shown in Fig. 16.59. This kind of relation is called a sine function. The voltage between the terminals of an outlet alters accordingly. It reaches its highest value, the peak voltage, twice per oscillation (120 times per second). It also takes the value of 0 V twice per oscillation.

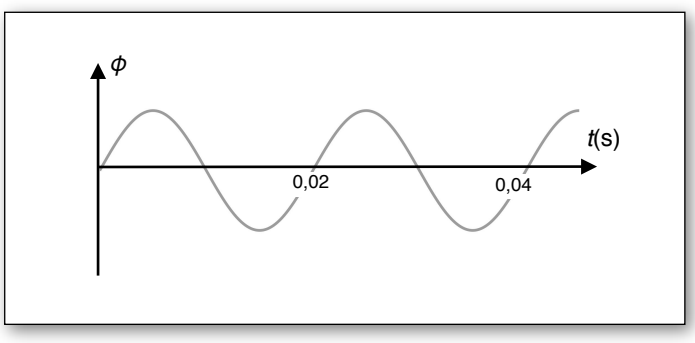

Fig. 16.59
The potential of the non-grounded terminal of the wall outlet as a function of time

We will now experiment a little with alternating voltages. We will need a sine wave generator for this. This is a power supply that produces an alternating voltage in the form of a sine wave whose frequency can be adjusted.

We connect a light bulb to the sine wave generator and set it so that the voltage makes one oscillation per second. We also set the peak voltage at 5 V. As expected, the light bulb turns on and off in quick succession, twice per second. When it reaches peak voltage, the light bulb glows as brightly as when attached to a 5 V direct voltage generator. At all other moments, it is less bright or doesn't glow at all.

We now raise the sine wave generator's frequency (always remembering to keep the peak voltage the same). The light bulb flickers faster and faster, finally reaching a frequency of 20 oscillations per second (20 Hertz) where it burns evenly. The filament is too slow to keep up with the quickly alternating voltage. However, the lamp does not glow as brightly as it would with a direct voltage of 5 V. It is receiving less energy on average than it would receive from a 5 V direct voltage power source.

For comparison, we now connect a second light bulb (identical to the first one) to a direct voltage power supply and adjust the voltage so that the second light bulb glows exactly as brightly as the first one.

We find that 3.5 V of direct voltage is necessary for this. A light bulb connected to an alternating voltage with a peak value of 5 V receives the same amount of energy as one connected to a 3.5 V direct voltage. Therefore, the alternating voltage generator is said to have an *rms voltage* of 3.5 V (the technical term rms stands for *root mean square*, but you need not know where this comes from).

The exact conversion factor between effective voltage and peak voltage is $\sqrt{2}$. Therefore

$$\text{peak voltage} = \sqrt{2} \cdot \text{rms voltage}$$

When one talks about voltage when referring to alternating voltage, one means the rms voltage. The 117 V of a wall outlet represents the rms voltage. The peak voltage of the outlet is

$117\ \text{V} \cdot \sqrt{2} \approx 165\ \text{V}$.

The voltage shown on an alternating voltmeter is also the rms voltage.

We haven't answered the most important question yet: What is all of this good for? Why is alternating voltage used so much? The answer is because there is a simple method of changing alternating voltages: with the transformer. A transformer raises or lowers voltages. It does so with hardly any loss of energy. However, it is limited to working with alternating voltages.

We will see later on how a transformer works and why it is desirable to raise and lower voltage.

16.11 The dangers of electric currents

Electric currents are dangerous, and everyone knows this. But what exactly is the danger? What should or shouldn't be done with them? What should one be aware of?

An electric current flowing through the human body has a damaging effect upon it. A current of 50 mA can be fatal.

An electric current can flow through our body only when it touches *two* points that are at *different* potentials. The sparrow on an electric wire is unharmed because it is touching only a single conductor.

It should not be assumed, though, that nothing bad will happen if you touch one terminal of an electric outlet without touching the other one. If the connection to the ground through the feet is a good conductor, then the second point of contact is established.

One of the two terminals of the outlet is indeed harmless because it is grounded. It is possible to touch this grounded terminal without anything happening. However, if both terminals of the outlet look the same we do not know which of them is dangerous and which is not.

> Do not touch both terminals of an outlet. Don't touch just one of them either.

Power supply voltage can be dangerous in a very different way, as well. If an electrical device has become damp or wet, the water can create a conducting connection between the hand and some conductor at 117 V. Therefore it is possible to receive an electric shock if a part of an electric appliance is touched that is usually insulated, say the plastic handle of a hair dryer.

Moisture is especially dangerous in electric devices because water creates a perfect contact to the body where we touch the appliance.

> Avoid moisture when working with electric appliances.

If an electric device has a metal casing or if other metal parts of it can be touched, this is another source of danger. The insulation of a conductor with a high potential could have a defect and the conductor might touch such a piece of metal. This piece of metal would also be at a high potential. In order to deal with this danger, the metal casing of the device is connected with by means of the *ground wire* to the third terminal of the socket, the one in the hole of the outlet. You remember, that this terminal is at ground potential. If a conductor at 117 V now comes in contact with the metal casing, there will be a short circuit and the main house fuse will interrupt the electric circuit.

We often handle other voltages than 117 V. Which voltages are the dangerous ones? Which values of voltage cause a current in the body that is harmful? This depends upon how the conductors, between which the voltage is applied, are touched.

If they are touched by the fingertips of the same hand, the danger is smaller. Because of the smaller contact surfaces, the current is small. In addition, the current is flowing through one hand, so that only this hand is affected.

However, if two conductors with different potentials are touched by each hand, the danger is greater. If the whole hand is the contact surface, the current is strong. Moreover, to get from one hand to the other one, the current flows, at least partly, through the heart—and this is especially dangerous. For this reason, it is advisable not to touch dangerous points at voltages above 40 V.

> Do not ever touch two conductors that have a voltage of more than 40 V between them.

Sometimes clothing can be charged with electricity. The voltage relative to ground can be several tens of kV (1 kV = 1000 V). In spite of this, touching them is not dangerous because clothing discharges so quickly that the current with the most dangerous values lasts only a microsecond (one millionth of a second) at the most and is not dangerous.

Exercises

1. A hair dryer that is not plugged in, falls into water. Why is it dangerous to use it after taking it out of the water?

2. The insulation of one of the conductors in the electric cable of a washing machine is worn away. The conductor comes in contact with the metal casing of the washing machine. When the machine is turned on, two things can happen. Explain.

17

Electricity and Energy

17.1 Electricity as an energy carrier

Light bulbs, electric motors, electric stoves, immersion heaters, and other electric appliances need energy. They get this energy with the energy carrier electricity. In most cases, the source of the energy is an electric power plant.

An energy current flows from a source to a receiver. A certain amount of energy (a certain number of Joules) flows per second. Remember that the energy E flowing past a given location in a time span t, divided by the time span t, is called the (strength of) energy current P. Expressed as a formula

$$P = \frac{E}{t}$$

The unit of P is

Joule/seconds = J/s.

The word "Watt" is used as the abbreviation for Joules per second. So

Watt = Joule/ seconds

Or

W = J/s.

In order to transport energy electrically (with the carrier electricity), a cable made up of two wires is needed. An electric current flows in these wires. It flows through one of them from the source to the receiver, and through the other one back to the source. There is an electric voltage between them, this means that they each have a different potential.

The strength of the energy current from the source to the receiver depends upon how strong the electric current in the wires is, as well as what the potential difference between the wires is. It is not difficult to derive this relationship.

We compare the arrangements in Figs. 17.1a and 17.1b. Both figures show a source-receiver-pair. Each setup has a power supply set at 12 V on the left, and a 12 V lamp on the right or two 12 V lamps on the right. The lamps are identical. This means that twice the energy current flows in arrangement Fig. 17.1b, compared to Fig. 17.1a. This is so because two lamps use twice the energy of one lamp. What are the voltage and electric currents in these cases?

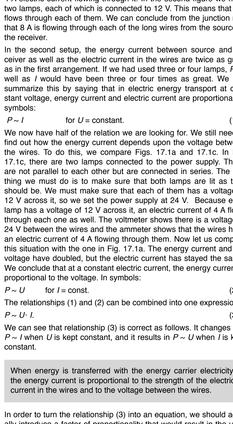

Fig. 17.1
The energy current is proportional to the electric current, if the voltage is constant. Compare (a) with (b). The energy current is proportional to the voltage if the electric current is constant. Compare (a) with (c).

The voltage is the same because the voltmeter shows 12 V in both cases.

What can we say about the electric current? The Ammeter in the first figure shows 4 A. This means that the lamp connected to 12 V, has a 4 A electric current flowing through it. In the second figure we have two lamps, each of which is connected to 12 V. This means that 4 A flows through each of them. We can conclude from the junction rule that 8 A is flowing through each of the long wires from the source to the receiver.

In the second setup, the energy current between source and receiver as well as the electric current in the wires are twice as great as in the first arrangement. If we had used three or four lamps, P as well as I would have been three or four times as great. We can summarize this by saying that in electric energy transport at constant voltage, energy current and electric current are proportional. In symbols:

$$P \sim I \qquad \text{for } U = \text{constant.} \qquad (1)$$

We now have half of the relation we are looking for. We still need to find out how the energy current depends upon the voltage between the wires. To do this, we compare Figs. 17.1a and 17.1c. In Fig. 17.1c, there are two lamps connected to the power supply. These are not parallel to each other but are connected in series. The first thing we must do is to make sure that both lamps are lit as they should be. We must make sure that each of them has a voltage of 12 V across it, so we set the power supply at 24 V. Because each lamp has a voltage of 12 V across it, an electric current of 4 A flows through each one as well. The voltmeter shows there is a voltage of 24 V between the wires and the ammeter shows that the wires have an electric current of 4 A flowing through them. Now let us compare this situation with the one in Fig. 17.1a. The energy current and the voltage have doubled, but the electric current has stayed the same. We conclude that at a constant electric current, the energy current is proportional to the voltage. In symbols:

$$P \sim U \qquad \text{for } I \text{ a const.} \qquad (2)$$

The relationships (1) and (2) can be combined into one expression:

$$P \sim U \cdot I. \qquad (3)$$

We can see that relationship (3) is correct as follows. It changes into $P \sim I$ when U is kept constant, and it results in $P \sim U$ when I is kept constant.

> When energy is transferred with the energy carrier electricity, the energy current is proportional to the strength of the electric current in the wires and to the voltage between the wires.

In order to turn the relationship (3) into an equation, we should actually introduce a factor of proportionality that would result in the units on the right agreeing with those on the left. We can therefore write:

$$P = k \cdot U \cdot I.$$

Fortunately, electric units have been defined so that $k = 1$, thus making it unnecessary to introduce any factor at all. We have:

$$P = U \cdot I \qquad (4)$$

If the voltage is given in Volts and the electric current in Amperes, the energy current obtained is in Watts.

Equation (4) shows that we have the following relation between units:

W = V · A.

Equation (4) is one of the most important formulas in electricity. If the values of two of the three quantities P, U, or I are known, the value of the third one can be calculated.

Example:

As soon as the electric current in a house rises above 16 A, the fuse interrupts it. What is the maximum amount of energy per second that can be obtained from the wall outlet?

Using $I = 15$ A and $U = 120$ V, the result according to formula (4) is

$P = 120 \text{ V} \cdot 15 \text{ A} = 1800$ W.

Two room heaters, each of which needs 1000 W, make the fuse disconnect.

Now we can understand what the printed information on an electric appliance means. The light bulb in Fig. 17.2 has "120 V/75 W" written on it.

Fig. 17.2
What does the inscription "120 V/75 W" mean?

First we will look at the part that says 120 V. Does this mean that the light bulb has a voltage of 120 V across it? Certainly not. Actually, this information is just a suggestion. It is recommended to use 120 V to light the bulb. No one can stop us from connecting the light bulb to a source with a smaller voltage. However, if we did so, the light would not be white but reddish. We could also connect the light bulb to a higher voltage, but that would reduce its lifetime.

What does the expression "75 W" mean? It means that an energy current of 75 W flows over the cable into the light bulb, assuming that the recommended voltage is used. Higher voltage means a stronger energy current and lower voltage means a weaker one.

Equation (4) also tells us how to measure the strength of the energy current that is conveyed with a two-core cable (if the energy carrier is electricity). First, one measures the electric current in one of the wires of the cable (it is the same as the one in the other wire), and then one measures the voltage across the wires. The product of the measured values is equal to the energy current.

There is an instrument called a wattmeter that can measure energy currents directly. A wattmeter has an inlet and an outlet for a two-core cable. The measurement is similar to methods for determining other currents. The wires of the two-core cable are cut in two, and the new ends are attached to the inlet and the outlet of the measuring device, Fig. 17.3.

Fig. 17.3
For measuring the energy current in a two core cable

Exercises

1. The headlight of a car is connected to the 12 V car battery. A 3.75 A electric current is flowing. What is the energy current flowing from the battery to the headlight?

2. The blinker of a car has 12 V/21 W printed on it. What is the electric current when the light is on?

3. Fig. 17.4a shows two lamps connected to a power supply. Use the values indicated by the three measuring devices to calculate
(a) The energy current coming out of the energy source;
(b) The energy current flowing to lamp L1;
(c) The energy current flowing to lamp L2.

4. Two motors connected in parallel are fed by a 12 V battery. An electric current of 2 A flows through the first motor and a current of 3 A flows through the second one.
(a) How much energy does the battery emit per second?
(b) What is the energy current flowing into motor 1? What is it for motor 2?

5. A 12 V battery and a 9 V battery are connected in series, and then both are connected to an electric motor. An electric current of 1.5 A flows. What is the energy current flowing to the motor? How many J does the 12 V battery emit per second? How many J does the 9 V battery emit per second?

6. Three monocells are combined to make one energy source, Fig. 17.4b. What is the voltage from A to B? There is an energy user (a load) attached to the connections A and B. There is an electric current of 10 mA flowing. Which of the three monocells drains first? How many Joules per second does the source emit? How many Joules do the monocells emit individually per second?

7. The energy source of a transistor radio is three monocell batteries connected in series. When the radio is turned on, an electric current of 60 mA flows on average. Each monocell has an energy content of 20 kJ.
(a) What is the energy current flowing out of the batteries?
(b) How long can a radio run on this set of batteries?

8. Make a list of the energy usage of various electric devices in your house. Where would it be worth saving energy?

9. An 80 V power supply is connected to a 2 kΩ resistor. What is the electric current flowing through the resistor? What is the energy current flowing to the resistor?

10. The resistor R in Fig. 17.5 has a resistance of 2 Ω. The ammeter shows 10 A. An energy current of 100 W is flowing into lamp L.
(a) What is the voltage produced by the battery?
(b) What is the electric current flowing through the lamp?
(c) What is the electric current flowing through the battery?

Fig. 17.4
(a) For Exercise **3**; (b) For Exercise **6**

Fig. 17.5
For Exercise **10**

17.2 Transmission resistance – energy loss in wires

A 6 V lamp is connected to a power supply by a very long cable, Fig. 17.6. At first, the voltage in the power supply is set to 0 V. We turn the voltage up until the voltmeter at the lamp reads 6 V. At this point the voltage is at its correct value and the lamp is burning as it should be. The ammeter shows that 5 A is flowing. The current is the same everywhere because we are dealing with a simple circuit with no branching.

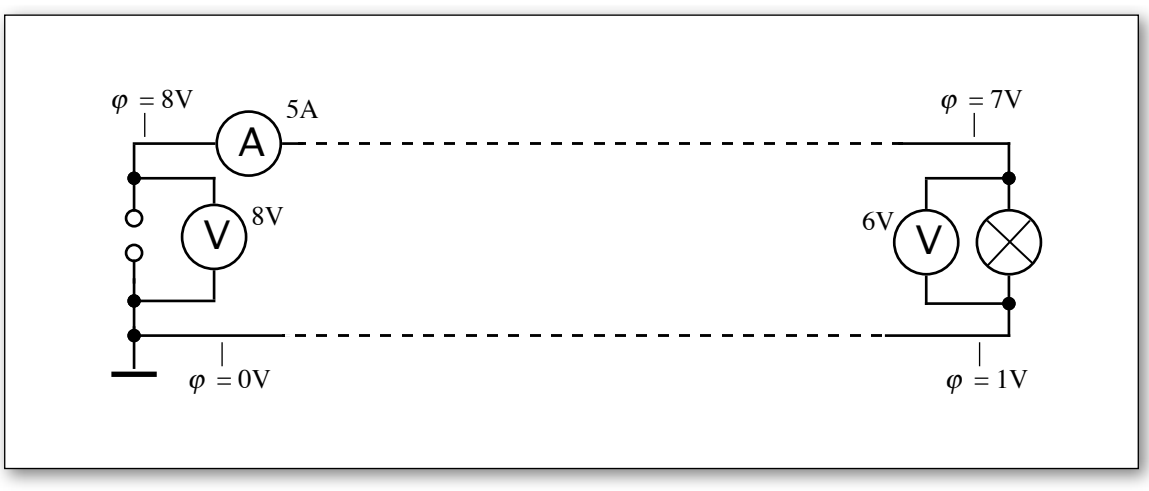

Fig. 17.6
The lamp's voltage is lower than that of the power supply. Two volts are needed for forcing the electricity through the long cables.

There is something odd here, though. The voltmeter on the left that shows the voltage between the terminals of the power supply is at 8 V. It would be expected to show 6 V. Why is this? We have always assumed that a wire would have the same potential everywhere. This assumption appears not to hold here. If it did, the potential of each of the two long wires would be the same on the left and on the right. Therefore the potential difference between the upper and the lower wire would also be the same on the right and left.

In the last chapter we learned about resistance in wires. We can now use this knowledge to explain this oddity. The current of 5 A which is flowing at every location in the circuit, not only needs a driving force to overcome the lamp's resistance, but it also needs one to flow through the wires. The two wires and the lamps must share the 8 V given by the power supply. As we know, the lamp needs a potential difference of 6 V. This leaves 2 V to drive the electricity through the wires. Since both wires, i. e. the forward and the return wire, of the cable are identical, each of them needs the same driving force. A voltage of 1 V is needed to drive the electric current through each wire. In other words: There is a potential difference of 1 V between the two ends of each wire. The potential values at four different locations of the circuit are given in Fig. 17.6

Transmission or conduction resistance is undesirable. Why? It is the reason for energy loss. It costs money.

We wish to calculate the energy loss in the circuit shown in Fig. 17.6. We can consider the circuit as three energy receivers connected in series. Fig. 17.7 shows the so-called equivalent circuit diagram. Both wires of the cable (the forward and the return wire) are replaced by a resistor: R_f and R_r. Now we can consider the lines drawn in Fig. 17.7 to be without resistance because the resistance is already taken into account by R_f and R_r. The potentials of the individual sections of the circuit are given in the Figure.

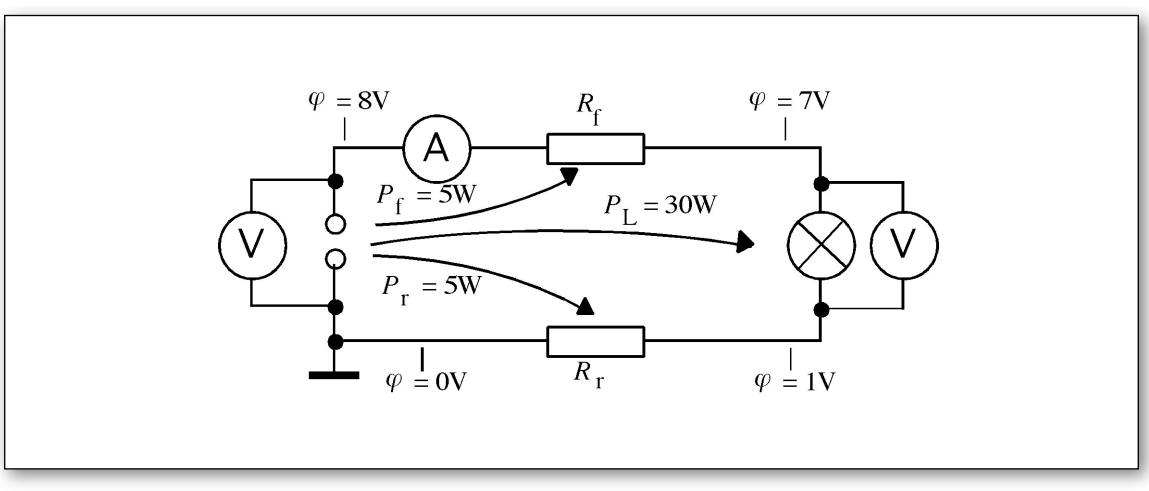

Fig. 17.7
Equivalent circuit diagram of the circuit in Fig. 17.6. The conductors have been replaced by resistor symbols.

There is an energy current flowing to each of the "devices" (the lamp, the resistor R_f and the resistor R_r). We calculate the three currents P_L, P_f and P_r using the equation $P = U \cdot I$.

The electric current is the same for each component, namely 5 A. The voltage between the two ends of the upper wire is

$U_f = 8\ V - 7\ V = 1\ V$,

And between the two ends of the lower wire is

$U_r = 1\ V - 0\ V = 1\ V$.

The voltage across the lamp is

$U_L = 7\ V - 1\ V = 6\ V$.

The resulting three energy currents are

$P_f = 1\ V \cdot 5\ A = 5\ W$

$P_r = 1\ V \cdot 5\ A = 5\ W$

$P_L = 6\ V \cdot 5\ A = 30\ W$.

The 5 W + 5 W = 10 W that is flowing into the resistors R_f and R_r cause those wires to become warm, and the 10 W are lost. This is a *lost energy current* of P_{loss}. Therefore

$P_{loss} = 10\ W$.

It is possible to use the equation

$R = U/I$

to easily calculate the transmission resistance. Using

$U_f = U_r = 1\ V$

and

$I = 5\ A$

we obtain

$R_f = R_r = 0.2\ \Omega$.

Each of the wires has a resistance of 0.2 Ω.

Exercises

1. A large motor is connected to a 200 V power supply by a long cable. Each of the two wires in the cable has a resistance of 0.5 Ω. An electric current of 8 A is flowing through the wires.

(a) What is the energy current leaving the power supply?
(b) What is the loss in the wires?
(c) How many Joules per second make it to the motor?

2. Fig. 17.8 shows two arrangements in which a lamp receives energy through a long cable from a power supply. Both are 60 W lamps. The lamp in 17.8a needs a voltage of 12 V. Therefore an electric current of 5 A is flowing. The lamp in 17.8b needs 24 V, so a current of 2.5 A flows. Each of the wires has a resistance of 1 Ω. Calculate the following for both arrangements:

– the voltage between the ends of a wire;
– the voltage at the terminals of the power supply;
– the energy loss.

Compare the losses in the two arrangements. Formulate a rule.

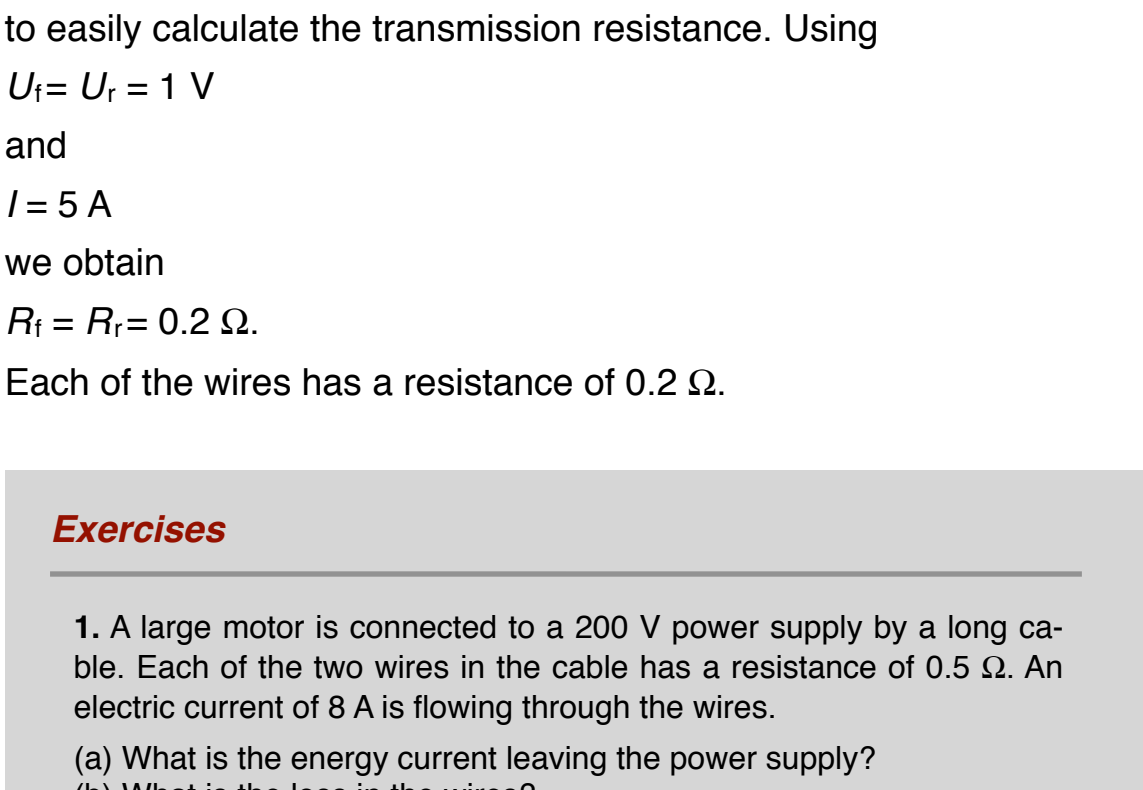

Fig. 17.8
In both cases, the same energy current reaches the lamps. The losses are different, though.

18

The magnetic field

18.1 Some simple experiments with magnets and nails

Magnets can attract or repel each other. Attraction or repulsion originates in their poles.

If a bar magnet with a pole at either end is hung horizontally from a thin thread so that it can rotate, it will orient itself north to south. One pole will point to the north and one will point to the south. This means that there are two types of poles. The pole that points to the north in our experiment is called the north pole. The other one is called the south pole.

Most magnets have only one north and only one south pole. However, there are magnets with more than two poles, for instance two north poles and two south poles. Magnets with only one pole, say only a north pole, do not exist. Fig. 18.1 shows three different magnets.

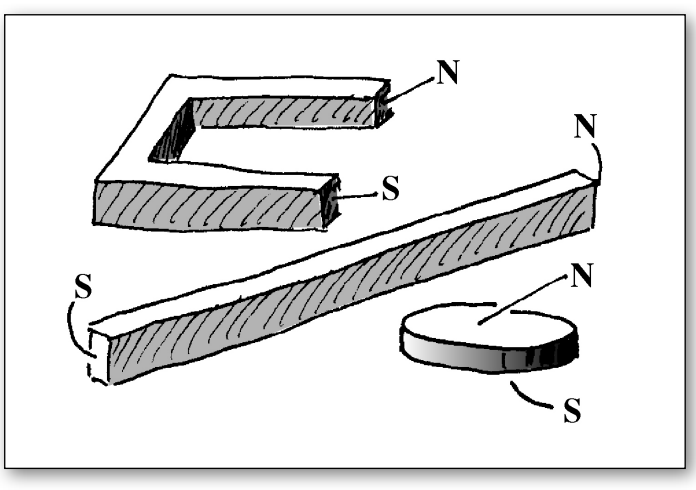

Fig. 18.1
Horseshoe magnet, bar magnet, and pill shaped magnet

Attraction between two magnets can only occur if a north pole faces a south pole. Two like poles repel each other, meaning two north poles or two south poles, Fig. 18.2. The closer the poles are to each other, the stronger the attraction or repulsion is.

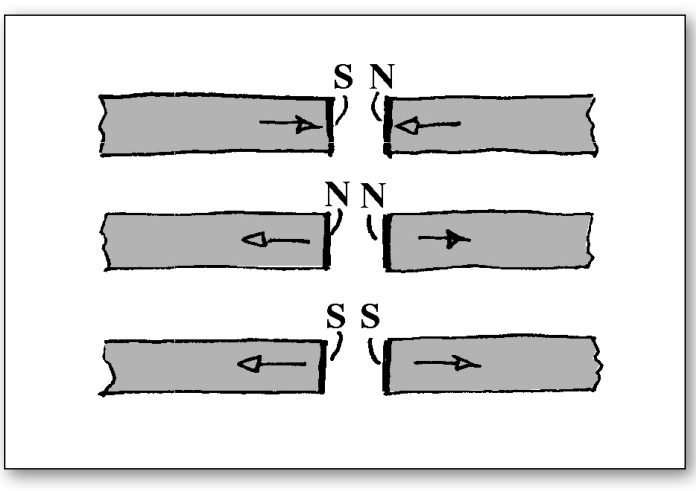

Fig. 18.2
Unlike poles attract each other, and like poles repel each other.

> Like poles repel each other, unlike poles attract each other.

The statements above have dealt only with how magnets react to each other. There is, however, another phenomenon that is closely related to what we have just discussed, but differs in one basic point, Fig. 18.3. A magnet attracts iron objects such as nails or paper clips. It always attracts iron and never repels it.

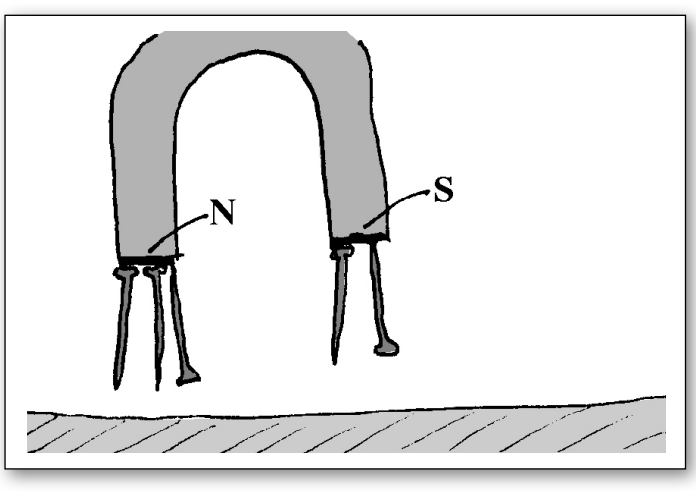

Fig. 18.3
Objects made of soft iron are always attracted to a magnet and never repelled by it.

This behavior does not disagree with the statement written above. Actually, it can be explained by it. The fact that a nail is attracted to a magnet has to do with different poles attracting each other. As soon as a nail is near a magnet, it becomes one itself. When a nail is put near the north pole of a magnet, Fig. 18.4, the end of the nail near the magnet becomes a south pole. In Fig. 18.4a, the nail's head becomes the south pole. The other end, the point, is then the north pole. It is clear how the nail is attracted to the magnet's north pole.

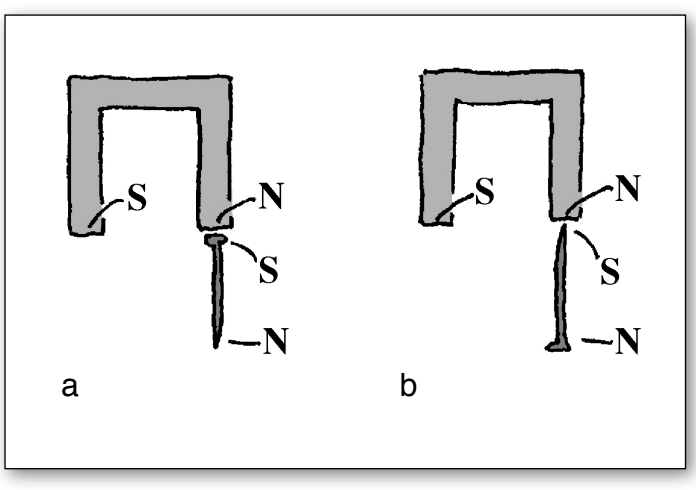

Fig. 18.4
A south pole forms at the end of the nail closest to the magnet's north pole.

If the nail is flipped over, Fig. 18.4b, a new south pole is created at the end of the nail pointing to the magnet (this time it is the point of the nail). The other end, the head, becomes the north pole.

When the nail is removed from the magnet, it loses its poles.

The phenomenon in Fig. 18.5 can now be understood. The second nail hangs from the newly created north pole of the first nail.

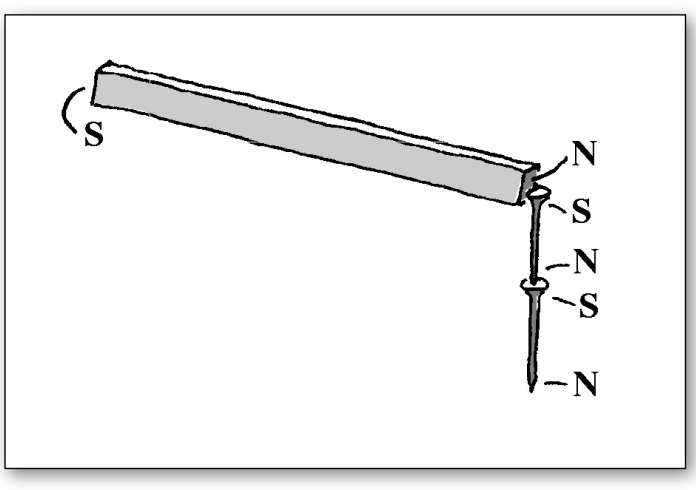

Fig. 18.5
The nail below hangs by its south pole from the north pole of the nail above it.

In order to emphasize that a "real" magnet does not lose its poles as easily as a nail, it is called a *permanent magnet*. A nail is not a permanent magnet.

The creation of magnetic poles on an object is called *magnetizing*. When we put the nail near the permanent magnet, we magnetized it. A permanent magnet is magnetized when it is produced in the factory.

Materials such as the iron nails are made of are called *magnetically soft.* Magnetically soft substances become magnetized when they are near a magnet. They are easily magnetized. They also lose their magnetism when moved away from a magnet.

The material that is used to make permanent magnets is called *magnetically hard*.

Even magnetically hard materials can lose their magnetism. A little "violence" is needed, however. If a permanent magnet is heated to about 800 °C, it will lose its magnetism. The magnetic poles disappear and never return even when the magnet is cooled down again. Try it yourself if you have a magnet you wouldn't mind giving up for this.

A typical magnetically soft material is soft iron. This is a type of iron that is used for making nails, for example. Magnetically hard materials have a complicated composition but they are mostly made up of iron.

There are also materials that fall somewhere between these two extremes. Steel would be a case of this. Steel also creates poles when it is put near a magnet but when it is removed, the poles do not disappear completely.

Steel can be permanently magnetized. Magnets can be made of it. You can try this yourself with a steel knitting needle. The magnetizing process is especially effective if you stroke one pole of a strong magnet several times in the same direction along the knitting needle.

This effect is used to store data on magnetic tapes, video cassettes, computer discs, and on credit cards. There is a very thin layer of a material on the data carrier (tape, band, etc.) that can be easily magnetized, and keeps its magnetism. The data is stored by magnetizing this layer along a line with a certain pattern.

Exercise

Someone claims that there are not only two, but four different types of magnetic poles. He gives you two magnets. One is the usual type with a north and a south pole. He also gives you another one that he says has an A pole and a B pole. What kind of experiments could you do in order to prove to him he is wrong?

18.2 Magnetic poles

Up until now, some of our statements about magnetic poles have been somewhat vague. Where exactly, are the poles located on a magnet? Where do they start and where do they end?

We will perform a very simple experiment with two strong, identical horseshoe magnets. First we will lift a heavy iron block with one of the magnets, Fig. 18.6.

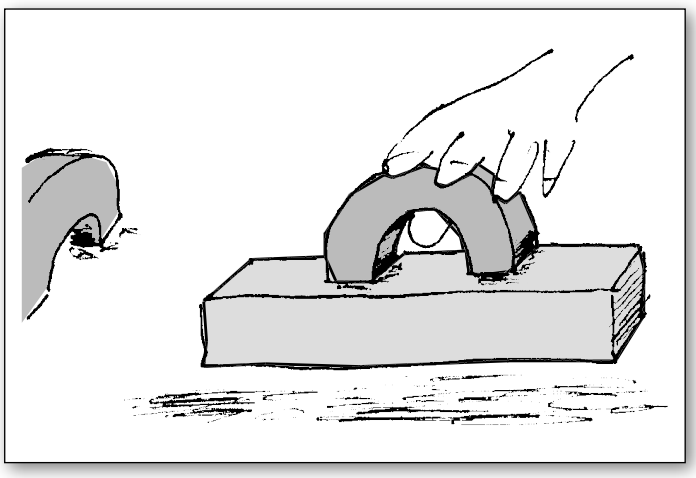

Fig. 18.6
The iron block can be lifted by just one of the horseshoe magnets.

Now we will put the two magnets together so that the south pole of one of them touches the north pole of the other and vice versa, Fig. 18.7. We will now attempt to use this ring shaped structure to lift our block. It doesn't work. The block does not hang from it. The poles' effect has disappeared. We can say that the poles themselves have disappeared, they have cancelled each other.

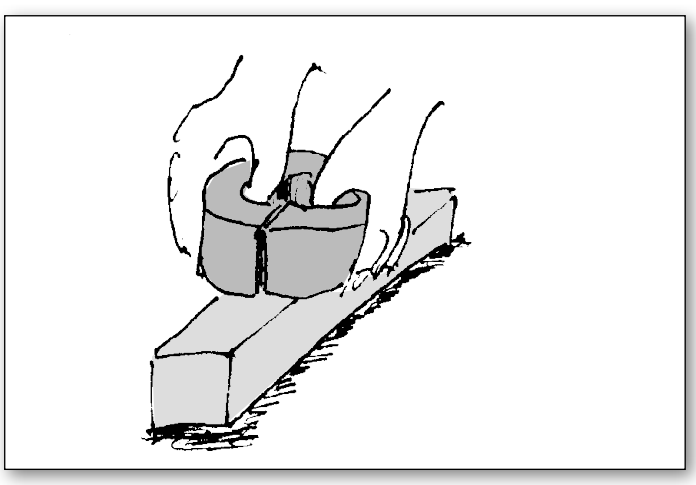

Fig. 18.7
The horseshoe magnets form a ring. The block of iron does not hang from this ring.

We wish to describe this observation more precisely. There is something at the magnetic poles of a magnet that we will call *magnetic charge*. This magnetic charge is located at the surface of the magnet. In the case of the horseshoe magnets on Fig. 18.6, it resides at the surfaces of the two ends.

Since the north and south pole charges cancel each other out, we can conclude that the magnetic charge has two different signs. It makes no difference which one of the poles is called positive or negative. It must only be agreed upon. It has been arbitrarily decided that the charge at the north pole is called positive and the one at the south pole is called negative.

If the same amounts of positive and negative magnetic charges are brought together, the resulting charge is zero. The positive and negative magnetic charges compensate each other. (It is the same as if you owed someone 100 Dollars and had 100 Dollars in the bank. The money you have would total 0 Dollars.)

The experiment in Fig. 18.7 can now be easily explained. The same amount of positive and negative magnetic charge is brought together at each of the two magnet's two contact surfaces. We can also draw another very simple conclusion from this: One magnet contains exactly as much positive as negative magnetic charge.

> A magnet contains exactly as much positive as negative magnetic charge.

This statement is valid for every magnet, even for a very asymmetrical one like in Fig. 18.8a. The magnetic charge is located at the two ends. Because the surface area of the north pole is larger than that of the south pole, the charge is more concentrated in the south pole. Fig. 18.8b shows another, more unusual magnet. It has a north pole but two south poles. Again, the charge at the north pole equals the sum of the charges at both the south poles.

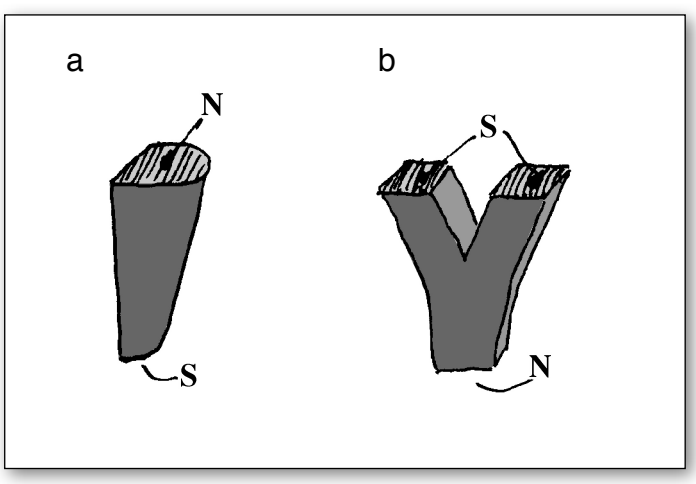

Fig. 18.8
Even for these unusually formed magnets, the north pole charge is equal to the south pole charge.

18.3 Lines of magnetization

We now find it easy to explain another phenomenon familiar to us. If a bar magnet is broken, two new magnetic poles are formed, Fig. 18.9. This breaking process can be repeated again and again. The result is always complete magnets, and each new piece has as much north pole charge as south pole charge.

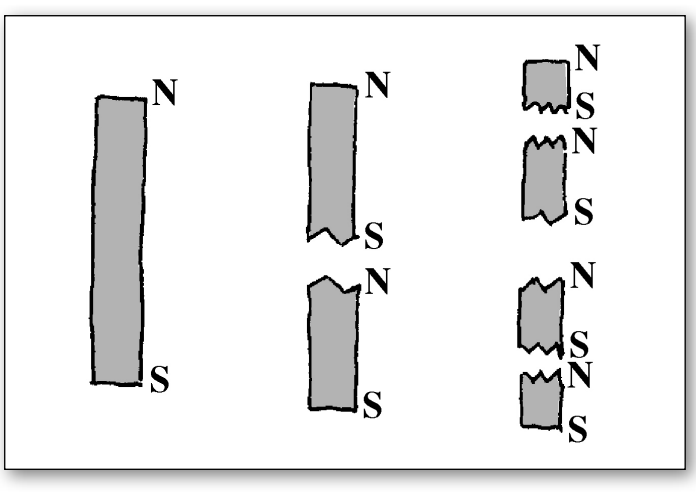

Fig. 18.9
If a bar magnet is broken, two new poles form at the surfaces of the breakage.

If a non-magnetized piece of steel is broken in half, there will be no magnetic charges at the breaking points. However, if a piece of magnetized steel is broken, new poles appear. We can conclude from this that magnetizing a piece of iron alters the entire piece and not just the locations of the poles.

Like every other substance, iron is made up of tiny particles called atoms. The atoms in iron are each magnetic, meaning that every atom is a tiny little magnet. However, as long as the iron has not been magnetized the atomic magnets are oriented irregularly. The result is that the piece of iron itself displays no magnetism, Fig. 18.10a. The effects of the individual little magnets cancel each other out.

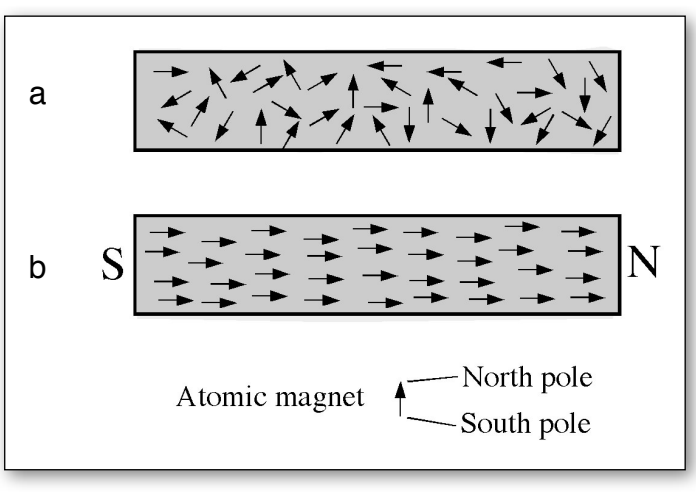

Fig. 18.10
(a) In a non-magnetized piece of iron, there is no order to the directions of the atomic magnets.
(b) In a magnetized piece of iron, the atomic magnets are aligned. On the left end surface of the magnet a south pole has formed, and on the right end, a north pole has formed.

In contrast, the atomic magnets in a permanent magnet or a magnetized piece of soft iron are all arranged regularly, Fig. 18.10b. This results in the left face surface of the magnet having only a south pole charge (negative magnetic charge) on it. The right surface then carries the north pole charge (positive magnetic charge).

Fig. 18.10b also shows how to graphically represent the magnetic state (magnetization) of an object. This representation can be done even more practically, though. Instead of showing the atomic magnets as many little arrows, continuous lines can be drawn. These are called *lines of magnetization*. They are drawn so that they show the direction the atomic magnets are oriented in. Each of these lines is given an arrow so that the line runs south to north, Fig. 18.11a.

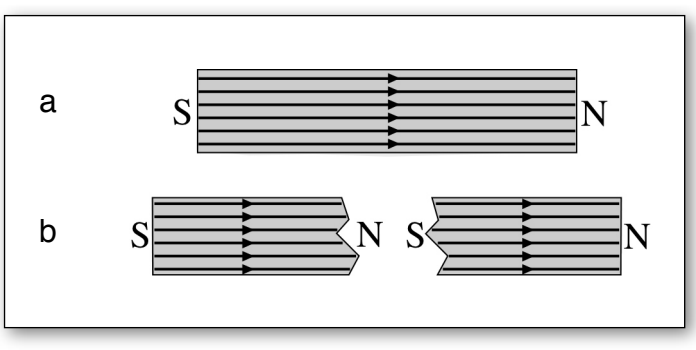

Fig. 18.11
(a) Graphic representation of the state of magnetization using magnetization lines.
(b) When the magnet is broken into two pieces, new poles form.

Lines of magnetization describe the magnetic state of a material. They begin at negative magnetic charges (south pole) and end at positive charges (north pole).

A sketch of lines of magnetization is significant. It tells us exactly where the magnetic charge of a magnet can be found. The place where the lines begin is where the negative magnetic charge is, and where they end is where the positive magnetic charge is. It also tells us what happens when a magnet is broken in two. If, for example, a magnet like the one in Fig. 18.11a, is broken in two, as in Fig. 18.11b, the magnetization lines end at the right side of the left hand piece. The new north pole is now there. Lines of magnetization begin at the left end of the piece of magnet on the right. A new south pole has been created there.

Iron can be magnetized in many different ways. We will now look at a somewhat unusual magnet, Fig. 18.12a. Fig. 18.12b shows how it is magnetized during production. The magnetization lines show us what happens when the magnet is broken into two pieces. Poles are only created at the upper breakages, and not at the lower ones, 18.12c.

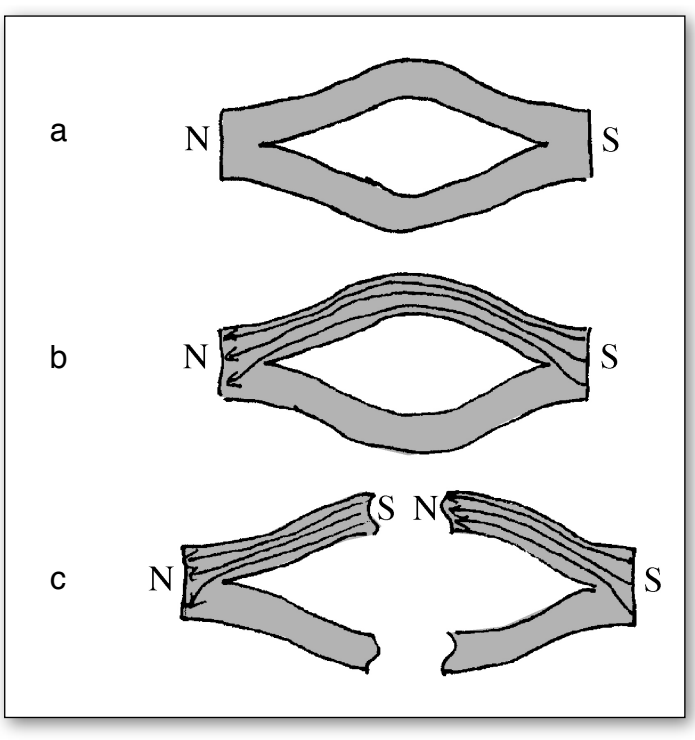

Fig. 18.12
(a) An unusual magnet.
(b) We cannot actually see that the magnetization lines run only through the upper section.
(c) When the magnet is broken, poles form only on the broken ends above and not on the ones below.

You see that the position of the poles can be determined from the lines of magnetization. Can this be done in reverse, though? Is it possible to draw the magnetization lines when the orientation of the poles is given? We consider Fig. 18.13a. The magnet has four poles, all on one side. What would the lines of magnetization look like in this case? It is easy to see that there are several solutions. Magnets with the kinds of poles found in Fig, 18.13a can be manufactured in different ways. Figs. 18.13b and 18.13c show two possibilities. It is not possible to see by looking at the magnet exactly what the magnetization is. One method of telling the possibilities apart would be to break the magnet in half. When the magnet in Fig. 18.13b is broken in half, no new poles appear, Fig. 18.13d. On the other hand, if the magnet in Fig. 18.13c is broken, a new north pole and south pole appear at each newly created breakage, Fig. 18.13e.

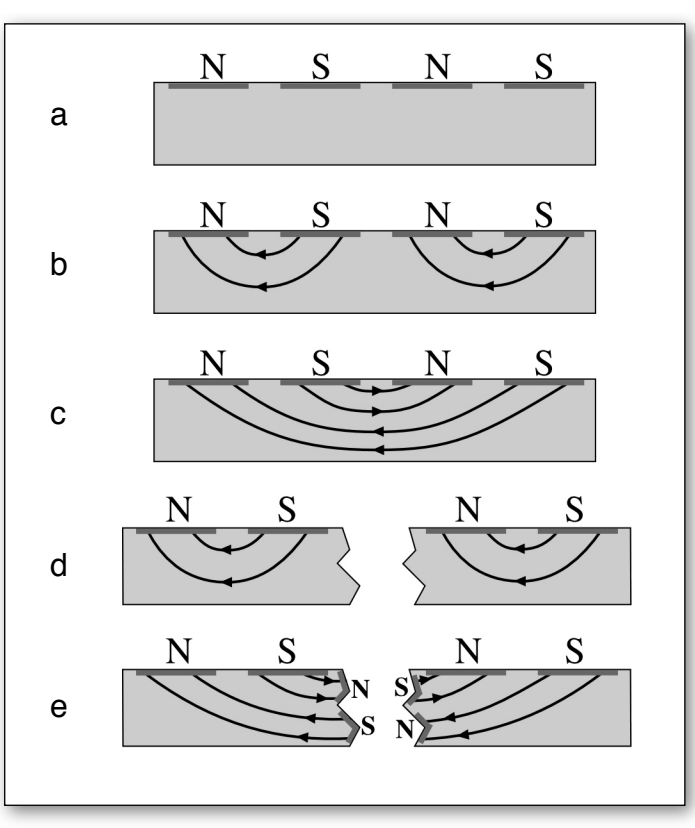

Fig. 18.13
(a) A magnet with four poles on its side.
(b, c) There are several possibilities for magnetization.
(d, e) The difference between b and c becomes clear when the magnets are broken through their middles.

Exercises

1. How would the magnetization lines of a horseshoe magnet look?

2. How might the magnetization lines in the magnet in Fig. 18.14a look?

3. How would the magnetization lines in the magnet in Fig. 18.14b look? Give two solutions.

4. A magnet is shaped like a small cylindrical disk. There are 3 north poles and 3 south poles on the lateral area of the cylinder. The north poles and south poles alternate and are evenly distributed over the circumference. How could the magnetization of the cylinder be? Give two solutions.

5. Someone gives you a steel ring and tells you that it is magnetized so that the magnetization lines run in a circle following the form of the ring. The magnet has no poles. How can you tell if this is true?

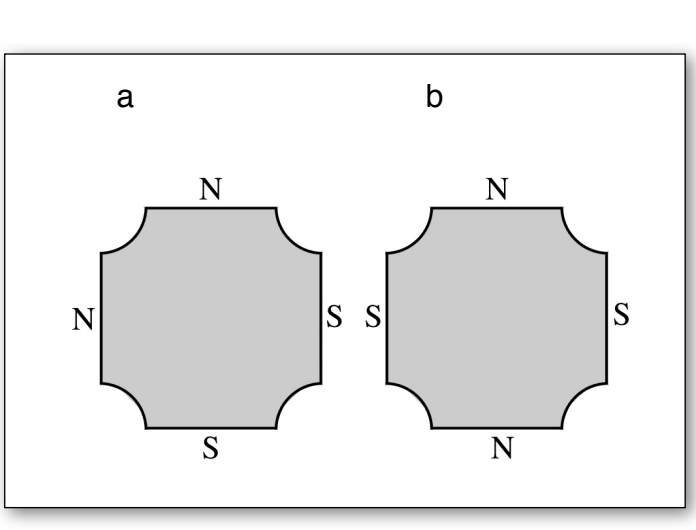

Fig. 18.14
For Exercises **2** and **3**. How would the magnetization lines run?

18.4 Magnetic fields

We will now take a look at a very different kind of problem. The two wagons in Fig. 18.15a are moving towards each other because the person is pulling on the rope. The wagons in Fig. 18.15b are being pushed apart by a spring. Fig. 18.15c shows two pistons in a cylinder. The piston on the left is pushed thereby, setting the one on the right in motion.

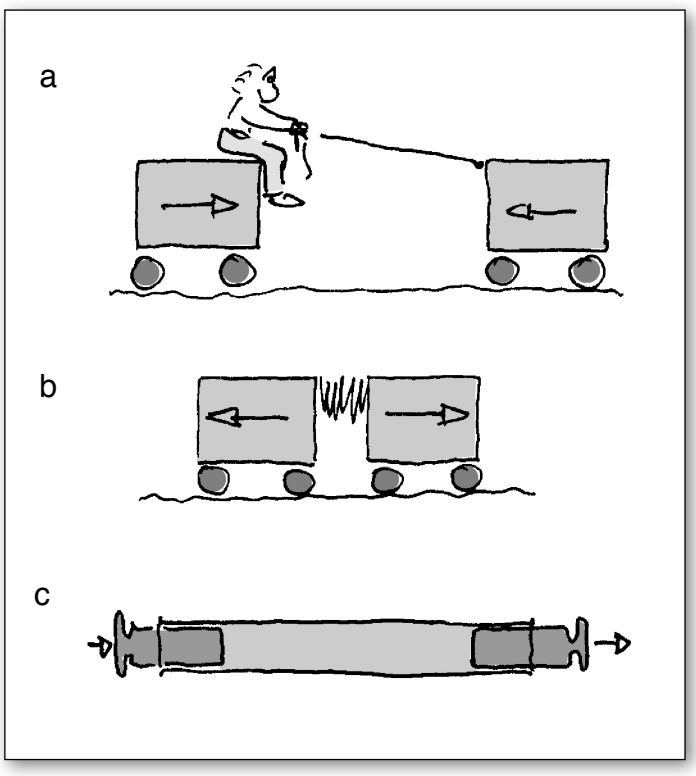

Fig. 18.15
The wagons in figure a are connected by a rope, and the wagons in b by a spring. The pistons in c are connected by air.

The three situations in Fig. 18.15 all have something in common. In each case, one object is set into motion at the cost of another one. One object receives momentum from another one.

What is important to notice here, is that if object A is to push or pull object B, there must be a connection between them. (A connection must exist for momentum to flow from A to B or from B to A.)

In Fig. 18.15, the first connection is a rope, the second one is a spring, and the third one is the air in the cylinder. In summary:

> If an object pushes another one away from itself or pulls another object towards itself, there must be a connection between them.

Now back to magnetism. There are magnets mounted upon two wagons, Fig. 18.16. Wagon A is moved toward wagon B, but before the wagons touch each other or the magnets come into contact, wagon B is set in motion.

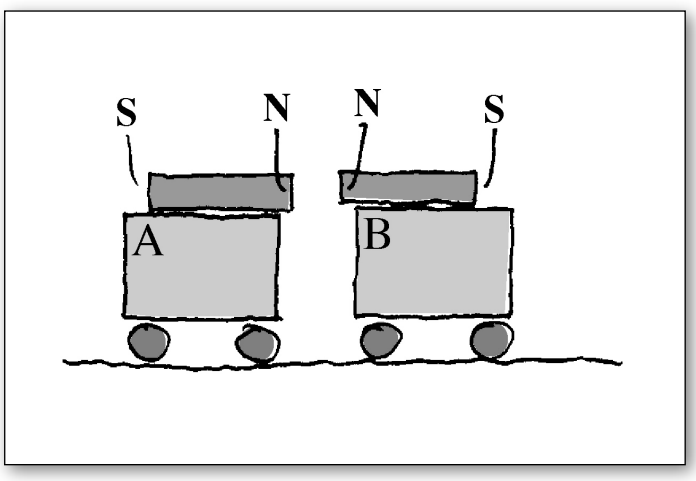

Fig. 18.16
The two magnets are connected by their magnetic field.

Of course, you say. We have already looked closely at this phenomenon. The north pole on the left repels the north pole on the right. However, if we take the statement above seriously, we can draw a new conclusion: There must be a connection in Fig. 18.16, through which the north pole on the left pushes the one on the right. It is clear that this connection is invisible, just as the air in Fig. 18.15c is invisible. The entity connecting the two north poles in Fig. 18.16 is called a *magnetic field*.

Take two very strong magnets and try pushing the two matching poles toward each other. You will feel the magnetic field trying to push the two poles apart.

There is a magnetic field attached to each of the poles of a magnet. If two poles of different magnets are brought together, the combined field between them acts like an elastic spring.

A magnetic field can push and pull just as a spring can. So there are not two kinds of fields. How does it happen that a magnetic field sometimes pulls and sometimes pushes? You will understand this better after reading the next section.

For the moment, the magnetic field helps us formulate an old rule more precisely. Previously we said “Like poles repel each other, unlike poles attract each other”. Take another look at the situation in Fig. 18.15b. Would you say that the two wagons repel each other? Of course not. It is better to say “the spring pushes the wagons away from each other.” We will now formulate a better rule about magnets:

> Like magnetic poles are pushed apart from each other by their field, unlike magnetic poles are pulled towards each other by their field.

18.5 Graphic representation of magnetic fields

The effects of a magnetic field surrounding a pole become weaker the further out it is from the pole. This is because the field is denser near the pole, and the density decreases with distance to the pole. This is similar to how air density decreases upwardly with distance to the Earth's surface. We cannot say that the field reaches to a sharply defined distance from the pole it surrounds. The field has no edge or sharply defined end, just as the air around the Earth has no strict boundary.

Now if we wish to represent the magnetic field in a drawing, we can express the various densities by drawing the field in dark gray near the poles and gradually make the gray tones lighter as we move away from the pole, Fig. 18.17.

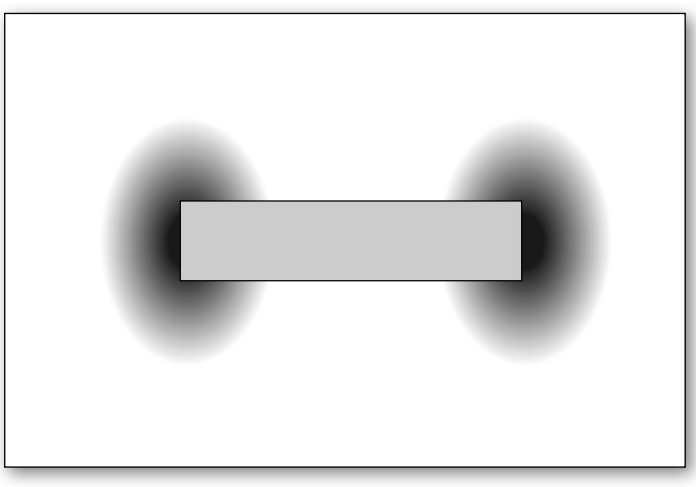

Fig. 18.17
Magnetic field densities are depicted by various shades of gray.

Another method would be to indicate the field with points. The points near the pole would be denser than further away from it, Fig. 18.18.

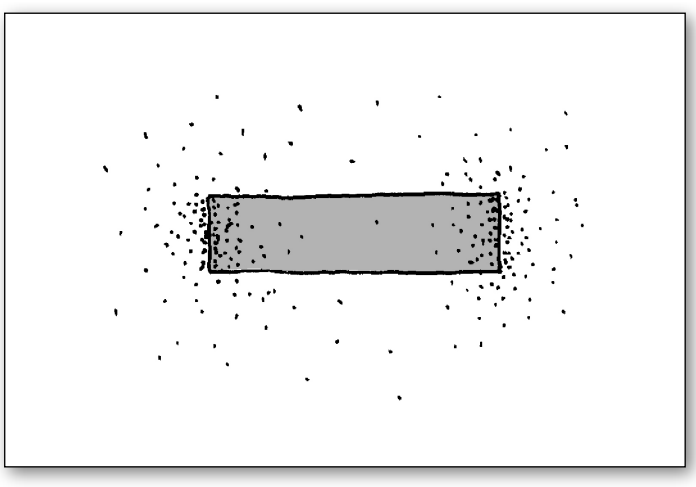

Fig. 18.18
The densities of the magnetic field are represented by point density.

This can be done better, though, and in order to do this we must first do a few experiments. First, we put a compass needle, i.e., a small magnet that can be rotated, near a magnet. The direction the needle will show depends upon where we put it. To every point in the field of our large magnet corresponds a certain direction. We can also say that at every point of a magnet the field has a certain direction.

It is not uncommon to attribute a certain direction to each point of a massive object. The fact that wood has a grain only means that every point of the wood has a direction that allows it to be easily split, Fig.18.19. We saw in the last section that there are prominent directions in iron when it has been magnetized.

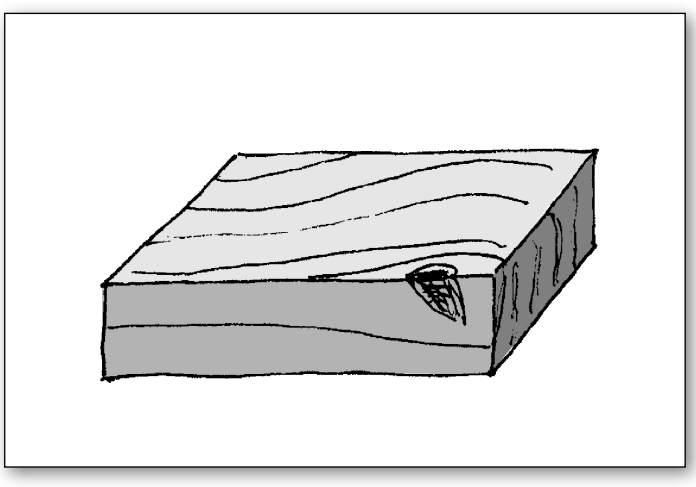

Fig. 18.19
The grain of wood tells us in which direction it is most easily split.

It is easy to make the individual directions of a magnetic field visible all at once. As an example, we ask what the directions are in the field of a bar magnet.

A plate made of a non-magnetic material (glass is best for this) is laid over the magnet. The glass does not change anything about the magnetic field. Iron filings are sprinkled on the glass plate, and the plate is tapped lightly. The filings create chains. These chains show the direction at every location of the field that a compass needle would also show. They show the field direction at every point.

Now back to our graphic representation of a field. We have seen that every "little piece" of the field has a certain direction. In order to show this direction in a drawing, we can proceed as shown in Fig.18.20. Instead of points (Fig. 18.18), we draw little arrows. The arrows are more numerous where the field is denser, and there are fewer arrows where the field is less dense. We draw the heads of the arrows at the end pointing away from the north pole of the big magnet.

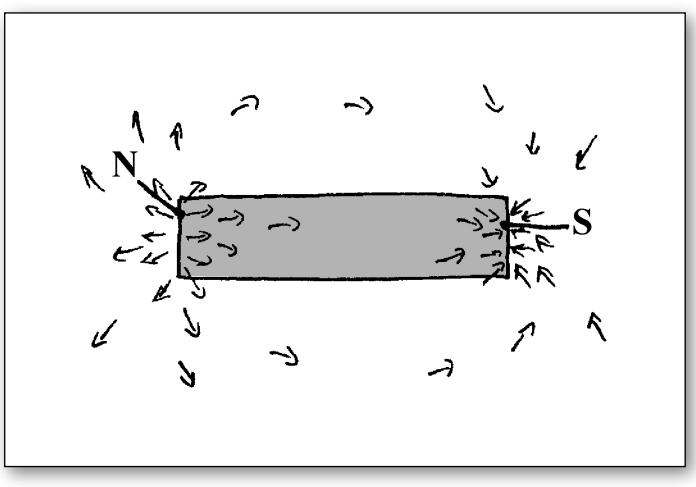

Fig. 18.20
The directions of the arrows show field directions, their density shows field density.

There is a more elegant method which uses so-called *field lines*, Fig. 18.21. Instead of the individual arrows in Fig. 18.20, one draws long continuous lines. The direction of the lines indicates the direction of the field. The lateral distance between the lines shows the field's density. If the lines are close together, the field is dense and if they are very far apart, the field is diluted. Arrows are added to the lines to show that the lines begin at the north pole and end at the south pole.

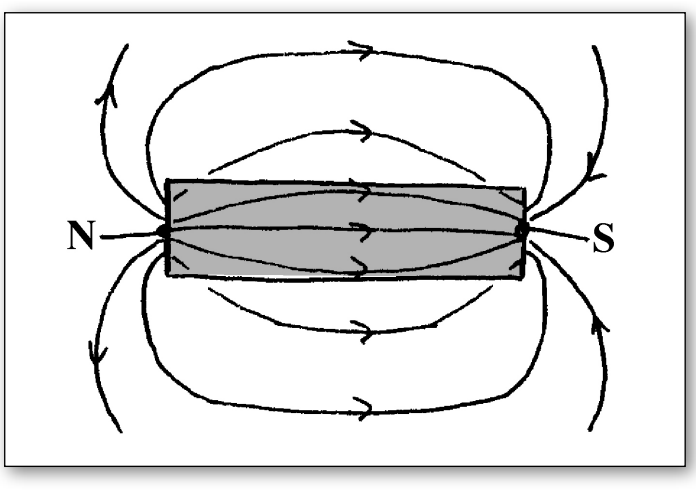

Fig. 18.21
Field lines. The closer together they are, the denser the field is.

In Fig. 18.21 you see that:

> Magnetic field lines begin at positive magnetic charges (north pole) and end at negative charges (south pole).

Don't mix up field lines with lines of magnetization. Both kinds of lines say something about directions. Magnetization lines tell us about the state of magnetized (visible) iron while field lines tell us about the state of an (invisible) field.

18.6 Magnetization lines and field lines

We have seen that it is possible to graphically represent the state of magnetization of materials as well as magnetic fields. We will now do both in one figure. We remind ourselves of the rules that magnetization lines begin at the south pole and end at the north pole, and that field lines begin at the north pole and end at the south pole. These rules can be summarized as follows:

> Magnetic field lines begin where magnetization lines end and vice versa.

It is a good idea to use different colors when representing both magnetization and field lines in one sketch.

We consider a magnet in the form of a ring from which a section has been removed. Fig. 18.22a shows the magnet with its poles. What would the magnetization lines look like? The simplest answer to this is shown in Fig. 18.22b. If a small compass needle is used to investigate the field, it shows the direction of the field lines, Fig. 18.22c.

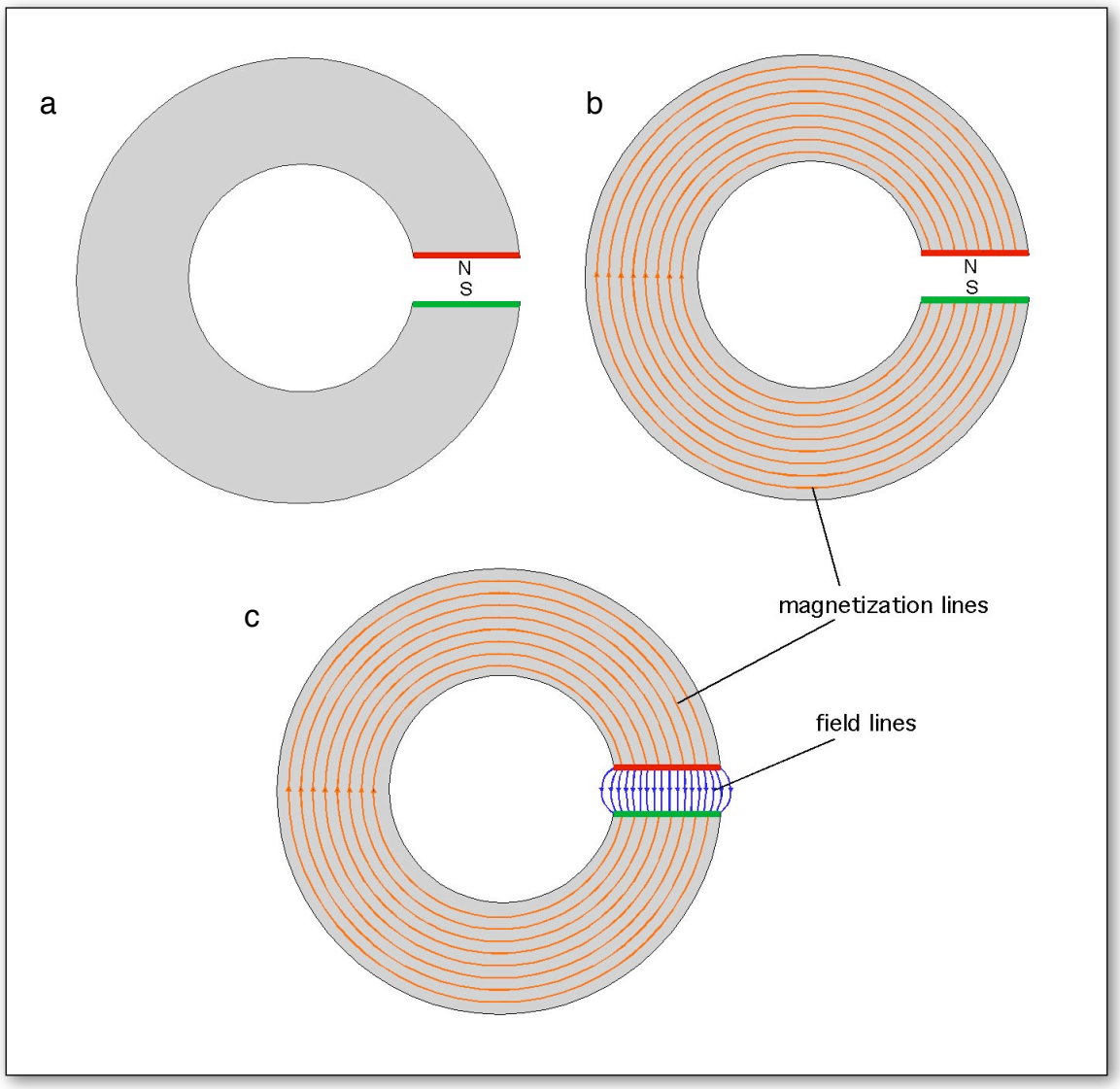

Fig. 18.22
(a) A ring magnet with poles.
(b) A ring magnet with magnetization lines.
(c) A ring magnet with magnetization lines and field lines.

Exercise

Fig. 18.23 shows an arrangement made up of a horseshoe magnet and a piece of soft iron. (a) Where do the poles develop on the soft iron? What kinds of poles (positive or negative) are these? (b) Where is the magnetic field? Sketch the field lines. (c) Draw the magnetization lines in the magnet and the soft iron.

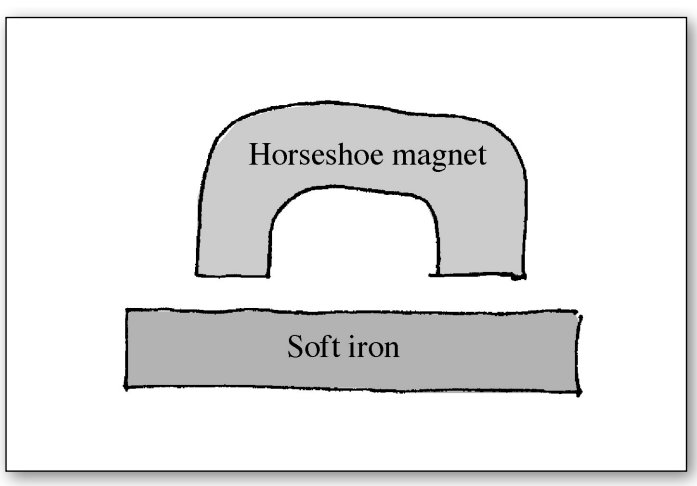

Fig. 18.23
For the exercise

18.7 Magnetic fields and matter

There is a fascinating aspect to magnets that we have not gone into yet. Magnets can push or pull through other bodies. We will now look more closely at this phenomenon.

A nail is hung from a thin thread and put near a strong magnet so that the nail is attracted to the magnet but does not touch it, Fig. 18.24. We now put plates of different materials in the space between the nail and the magnet.

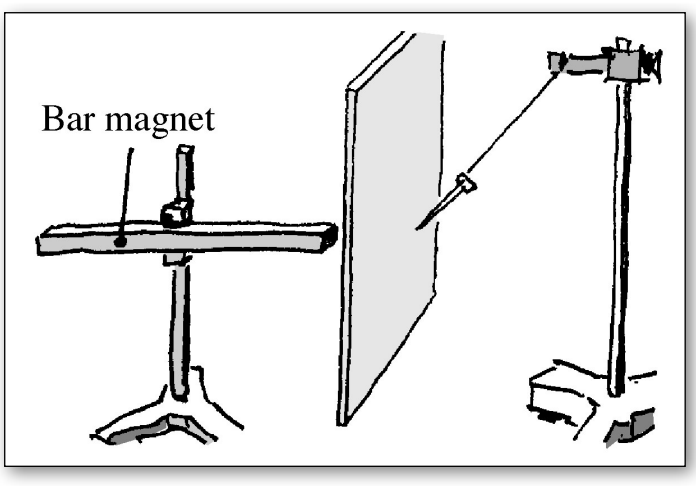

Fig. 18.24
Plates made of different materials are placed in the space between the magnetic pole and the nail.

In most cases, the nail stays where it is and doesn't seem to "notice" the plates at all. This is the case with plates of cardboard, wood, glass, and various plastics, as well as most metals such as aluminum, copper and lead. These are exactly the materials that are not attracted by magnets and cannot be magnetized.

A magnetic field just goes through these materials. It penetrates these materials as if they weren't there. Is this a surprise to you? Could it be that two things can exist in the same location in space, namely a material and a field? Actually, this shouldn't really be surprising. We all know another everyday example of when two "substances" share the same place: When light passes through glass, both light and glass are in the same place. In just this way, magnetic fields and copper, for example, can share the same place.

This situation changes, though, if we put a soft magnetic material, say an iron plate, between the nail and the magnet (see Fig. 18.24). The magnet releases the nail and it hangs down.

The iron plate does not allow the magnetic field through. We can formulate this more precisely if we first do a little experiment. We slide a thin iron plate between the magnet and the nail. The nail falls down. We then slide two thin iron plates in and then three, etc. Of course the nail falls every time. We can consider the two, three, or more plates as one thicker plate, Fig. 18.25. Now, not only does the field not come out on the right hand side of the thick plate, but the field never penetrates the plate, at least no deeper than the depth of the first thin plate. We have made another important discovery here:

The fact that soft magnetic bodies form poles when put into a magnetic field means that they react to magnetic fields by magnetizing.

A magnetic field does not penetrate far into soft magnetic substances.

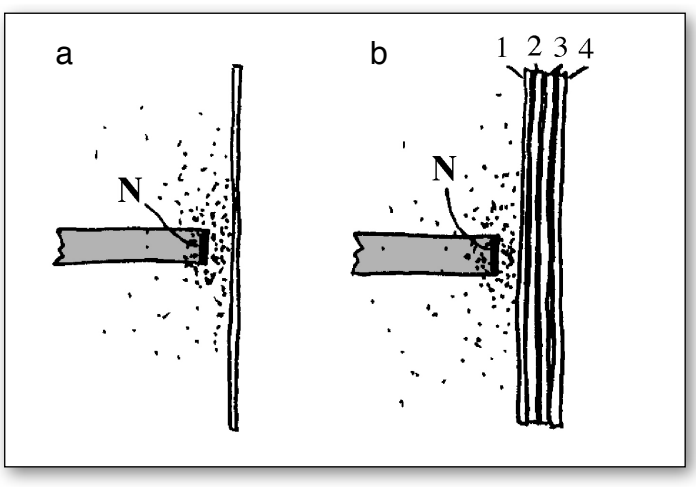

Fig. 18.25
The field does not go through the thin iron plate in figure a. It therefore cannot penetrate plates 2, 3, and 4 in figure b.

We want to investigate where the poles form on our plate. A "two dimensional" arrangement is best used here, Fig. 18.26, so that the fields can be made visible by iron filings. We find that a south pole forms on the rod of soft iron near the north pole of the magnet. A distance away from this south pole, and on both sides of it, a north pole forms. The magnetic charges at these two north poles are more strongly diluted than that of the south pole in the middle.

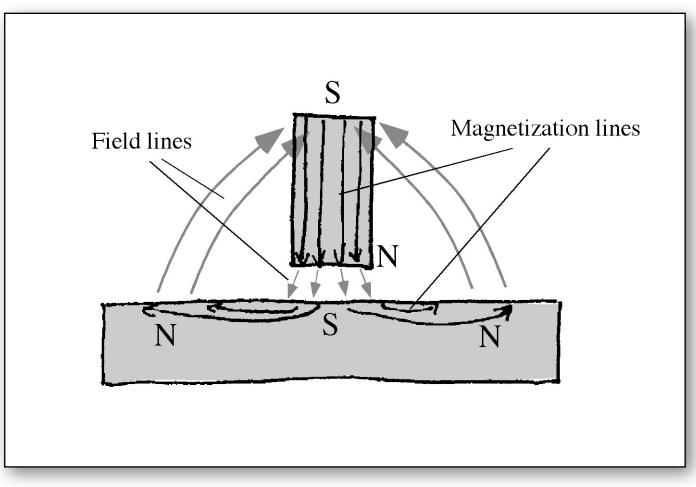

Fig. 18.26
The soft iron's north pole is made up of two parts that are separated by the south pole.

What does the pole distribution look like in the original arrangement where we held up an extended plate to the magnet and not a rod? The south pole forms directly in front of the magnet's north pole. The north pole of the plate of soft iron plate is, in this case, ring shaped, Fig. 18.27.

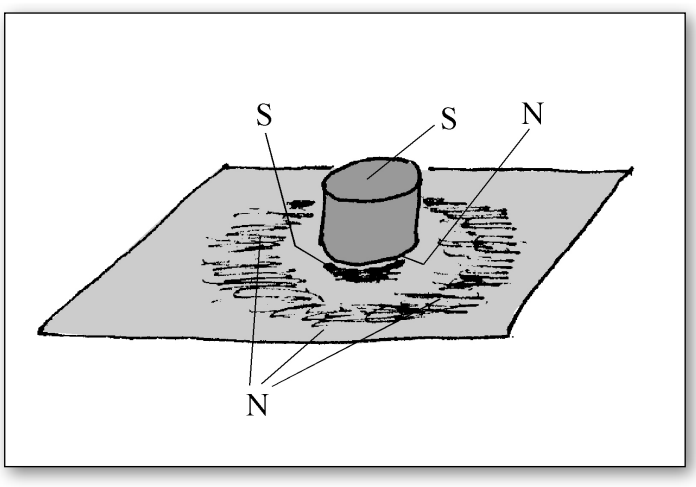

Fig. 18.27
The north pole of the soft iron plate forms a ring around the south pole.

Exercise

Draw the magnetization lines and the field lines for the magnet in Fig. 18.28a. A small plate of soft iron is put in the middle of the area covered by the field, Fig. 18.28b. What do the magnetization lines and field lines look like now?

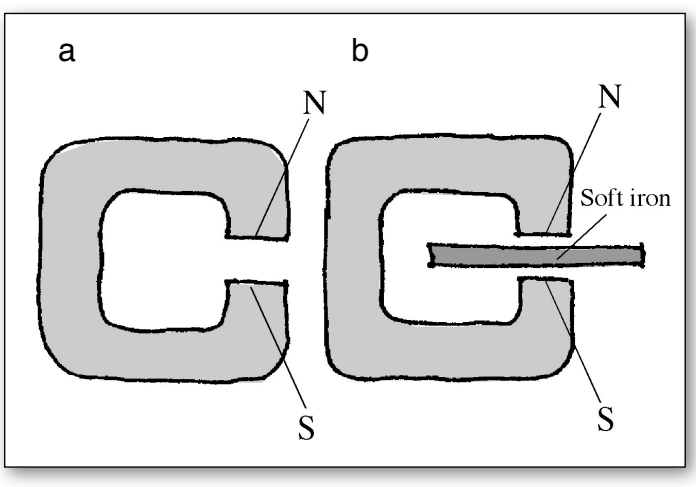

Fig. 18.28
How do the magnetization and field lines run before and after the soft iron plate has been inserted?

18.8 Energy of a magnetic field

In order to pull apart two strong magnets that are attached to each other, energy must be expended. Where does this energy go?

There is a strong magnet attached to a table with another strong magnet lying near it. The first magnet pulls the other one towards itself. The moving magnet can drive something for a short moment, a dynamo for instance. Energy is needed here as well. Where does this energy come from?

Compare the two situations. In the first one, energy (from the person trying to separate them) is used and a magnetic field is created. In the second case, energy is given off (to the dynamo) and the magnetic field disappears. We conclude that energy is contained in a magnetic field.

A magnetic field contains energy.

18.9 Electric currents and magnetic fields

We have a long wire like in Fig. 18.29. The wire can be connected to a car battery so that an electric current flows through it. It flows downward in the part on the right and upward in the part on the left. The circuit can be closed for only short periods because the wire's resistance is very small and the current is more than 50 A. One end of the wire is firmly connected to the battery. While observing the piece of wire, we touch the other battery terminal shortly with the other end of it. The pieces of wire jump apart. Something pushed them away from each other.

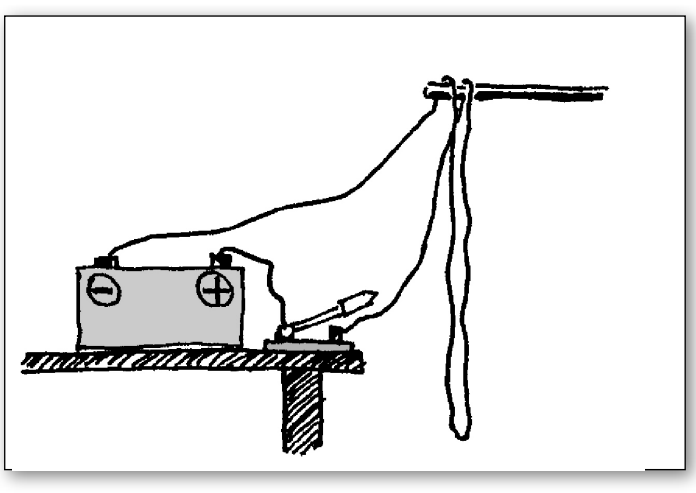

Fig. 18.29
When closing the circuit, the wires jump apart.

We repeat the experiment but shift the wires so that the current flows in the same direction in both vertically hanging sections, Fig. 18.30. This time the two sections of wire jump toward each other when the circuit is closed.

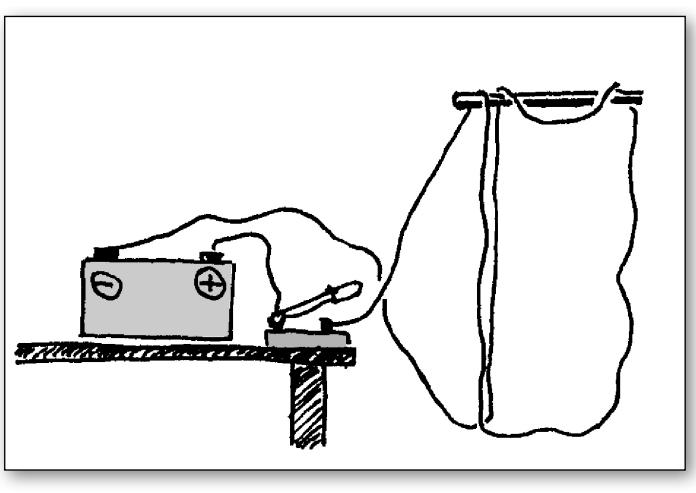

Fig. 18.30
When closing the circuit, the wires jump together.

What is the connection through which one wire attracts or repels the other?

The answer is easy to find. We put a compass needle near a wire in which a strong electric current can flow. As soon as the current is turned on, the needle points in a certain direction, Fig. 18.31. When the current is turned off again, the needle returns to its original position. Obviously, the wire is surrounded by a magnetic field as long as an electric current is flowing through it.

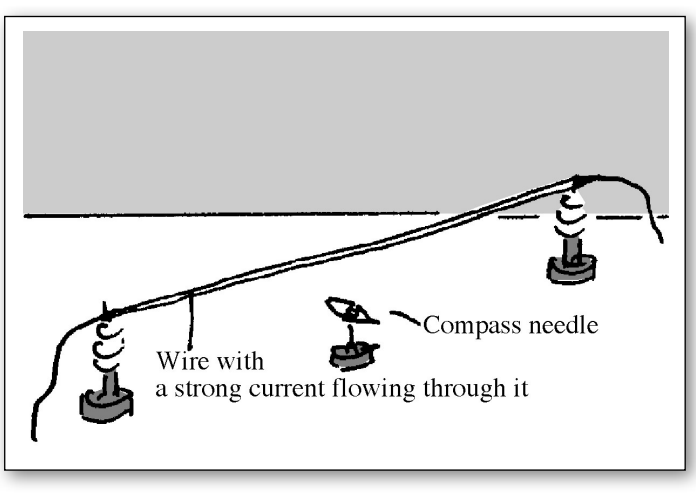

Fig. 18.31
As soon as the electric current is turned on, the compass needle changes its direction.

We will now investigate the direction and density of the field by moving the compass needle around a vertically hanging wire. It shows that the field direction is perpendicular everywhere to the direction of the wire. Moreover, every arrow showing the field direction lies upon a circle whose center is on the wire's axis.

The density of the field decreases with distance from the center, meaning as we move away from the wire. Fig. 18.32a shows the field represented by dots, Fig. 18.32b shows it with arrows, and Fig. 18.32c represents it with field lines. We see that the field is not attached to the wire by heads or ends of arrows, but with the sides of the arrows. In this case, we do not need to look for magnetic poles because they cannot exist here.

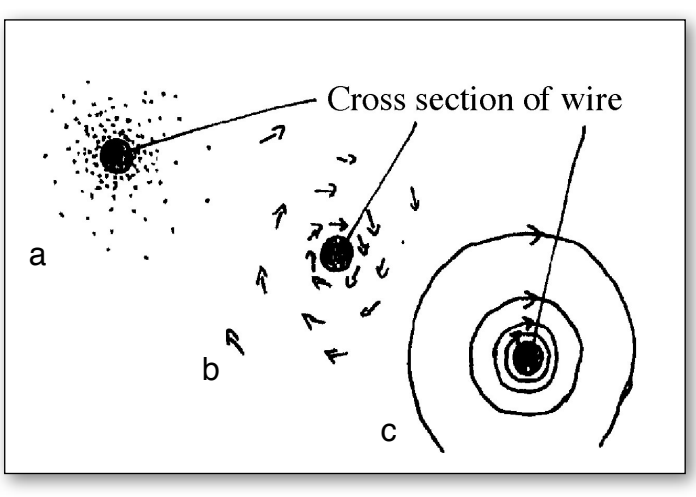

Fig. 18.32
Three different ways to represent the magnetic field around an electric current.

It would have been very odd in any case if we had found poles on the wire because it is made of copper which cannot be magnetized.

An electric current is surrounded by a magnetic field. If the electric currents in two parallel wires flow in the same direction, the magnetic field pulls the two wires together. If the currents are flowing in opposite directions, the field pushes the two wires apart.

The attraction and repulsion of electric currents has found very important technical applications. However, before this was possible, these attraction and repulsion effects had to be made stronger. If two hanging wires are just barely set in motion by a current of 50 A, the effort needed is simply too much compared to the effect.

There is a trick that can be used for making the magnetic field caused by the electric currents much denser. The wire is formed into a *coil* which allows it to pass by the same location several times.

In Fig. 18.33, the wire is coiled 100 times. Now, if a current of 1 A flows through it, a total current of 100 A flows through the cross section shown. As a result, near this bundle of wire we now have a magnetic field that is as dense as it would be if one wire had 100 A flowing through it.

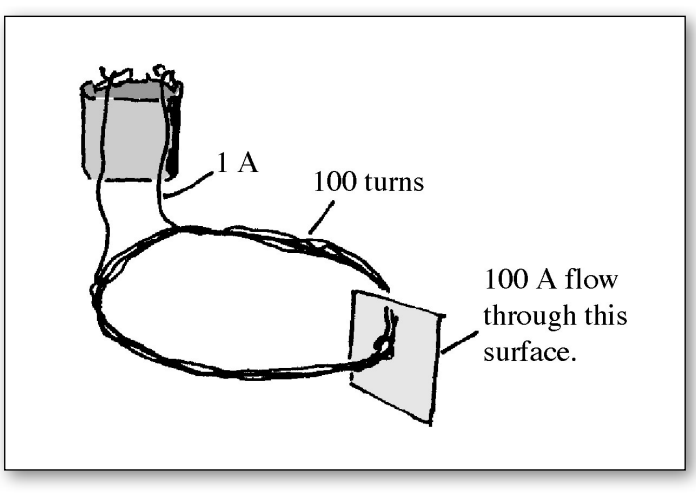

Fig. 18.33
The bundle of wires has the same field density as that of a single wire with a current of 100 A flowing in it.

A cylindrically formed coil is an especially useful arrangement (this is called a solenoid). In this case, the wire is wound cylindrically in many layers, Fig. 18.34. (Of course the wire must be insulated, otherwise the current might seek a short cut.)

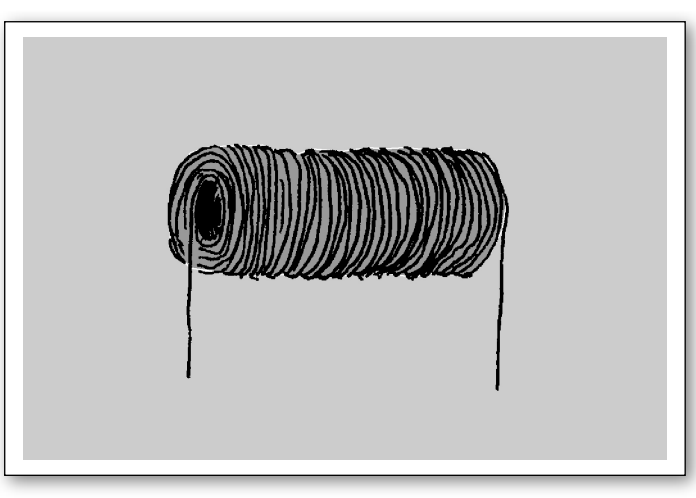

Fig. 18.34
Cylindrically formed solenoid

We investigate the direction and density of the field of a solenoid, for example, with the help of iron filings. The result is shown in Fig. 18.35. The field is densest inside the coil. There its direction is that of the cylinder's axis at every point.

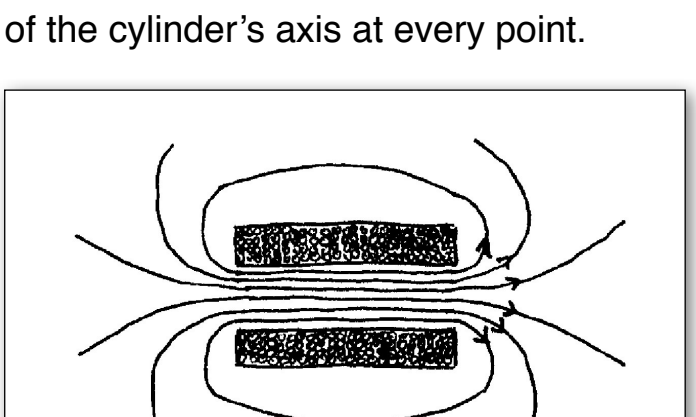

Fig. 18.35
Field of a cylindrically formed solenoid

Exercises

1. How would a coil be made so that no magnetic field is produced in its surroundings when an electric current flows through it?

2. In what direction does the magnetic field of a coil press upon its wires? How does the pressure depend upon the direction of the electric current in the coil?

18.10 Electromagnets

A magnet is mounted upon a small wagon. The wagon is put in front of a solenoid, Fig. 18.36. When the electric circuit of the solenoid is closed, the wagon with the magnet is pulled toward the solenoid, or it is pushed away – depending upon the direction of the electric current in the solenoid.

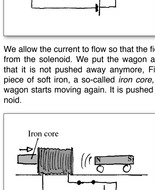

Fig. 18.36
The field pushes the magnet to the right.

We allow the current to flow so that the field pushes the wagon away from the solenoid. We put the wagon at a distance far enough so that it is not pushed away anymore, Fig. 18.37. Now we insert a piece of soft iron, a so-called *iron core*, into the solenoid. Now the wagon starts moving again. It is pushed further away from the solenoid.

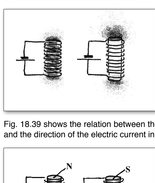

Fig. 18.37
If an iron core is inserted into the solenoid, the magnet moves even further to the right.

How is this possible? We remember what happens when a piece of soft iron is put into a magnetic field:

– The soft iron magnetizes, forming poles;
– The field is displaced.

In this case, the field is pushed out of the solenoid. It has the greatest density at the ends of the iron core, Fig. 18.38. The poles form on the end surfaces of the iron core.

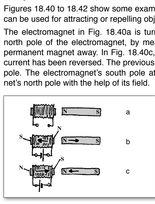

Fig. 18.38
The magnetic field is forced out of the coil by the iron core.

Fig. 18.39 shows the relation between the direction of magnetization and the direction of the electric current in the solenoid.

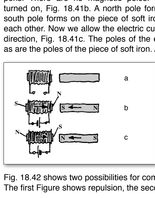

Fig. 18.39
If the direction of the electric current in the coil is reversed, the poles of the magnet are interchanged.

We have transformed a piece of soft iron into a magnet. The soft iron and the solenoid together make an *electromagnet*.

Electromagnets have an advantage over permanent magnets in that they can be turned on and off. They can also be adjusted to be strong or weak and their poles can be reversed.

Figures 18.40 to 18.42 show some examples of how electromagnets can be used for attracting or repelling objects.

The electromagnet in Fig. 18.40a is turned off. In Fig. 18.40b, the north pole of the electromagnet, by means of its field, pushes the permanent magnet away. In Fig. 18.40c, the direction of the electric current has been reversed. The previous north pole is now the south pole. The electromagnet's south pole attracts the permanent magnet's north pole with the help of its field.

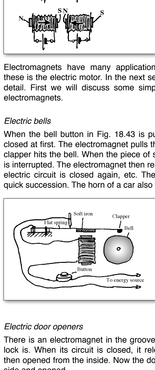

Fig. 18.40
Electromagnet and permanent magnet

In Fig. 18.41, a piece of soft iron is put in front of the electromagnet. As long as the electromagnet is turned off, Fig. 18.41a, nothing happens. There are no magnetic poles. Now the electromagnet is turned on, Fig. 18.41b. A north pole forms on its right end while a south pole forms on the piece of soft iron. The poles move toward each other. Now we allow the electric current to flow in the opposite direction, Fig. 18.41c. The poles of the electromagnet are reversed as are the poles of the piece of soft iron. Again, we have attraction.

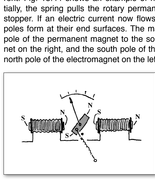

Fig. 18.41
Electromagnet and soft iron

Fig. 18.42 shows two possibilities for combining two electromagnets. The first Figure shows repulsion, the second, attraction.

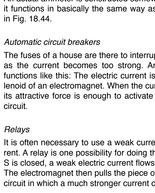

Fig. 18.42
Electromagnet and soft iron

Electromagnets have many applications. The most important of these is the electric motor. In the next section, we will go into this in detail. First we will discuss some simpler devices that work with electromagnets.

Electric bells

When the bell button in Fig. 18.43 is pushed, the electric circuit is closed at first. The electromagnet pulls the piece of soft iron and the clapper hits the bell. When the piece of soft iron is pulled, the circuit is interrupted. The electromagnet then releases the soft iron, and the electric circuit is closed again, etc. The clapper strikes the bell in quick succession. The horn of a car also works similarly to this.

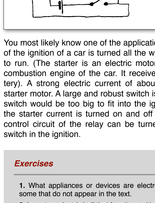

Fig. 18.43
An electric bell

Electric door openers

There is an electromagnet in the groove of a door frame where the lock is. When its circuit is closed, it releases the lock that can be then opened from the inside. Now the door can be pushed from outside and opened.

Electric clocks

There is an electromagnet in the kinds of electric clocks that do not have liquid crystal displays. A short current surge goes through the electromagnet at regular time intervals of once per second. The electromagnet causes the clock's second hand to move a little further with each current surge.

Ammeter

An electromagnet's strength depends upon the electric current flowing through it. This can be used in order to measure an electric current. Fig. 18.44 shows an example of how an ammeter works. Initially, the spring pulls the rotary permanent magnet right up to the stopper. If an electric current now flows through the two magnets, poles form at their end surfaces. The magnetic field pulls the north pole of the permanent magnet to the south pole of the electromagnet on the right, and the south pole of the permanent magnet to the north pole of the electromagnet on the left.

Fig. 18.44
How an ammeter works

The stronger the electric current, the greater the charges at the poles of the electromagnet, and the stronger the field's pull. The stronger the field pulls, the further the permanent magnet and its needle turn.

An actual ammeter is constructed somewhat differently from this, but it functions in basically the same way as the primitive one we used in Fig. 18.44.

Automatic circuit breakers

The fuses of a house are there to interrupt an electric circuit as soon as the current becomes too strong. An automatic circuit breaker functions like this: The electric current is conducted through the solenoid of an electromagnet. When the current reaches a given value, its attractive force is enough to activate a switch that interrupts the circuit.

Relays

It is often necessary to use a weak current to control a stronger current. A relay is one possibility for doing this, Fig. 18.45. When switch S is closed, a weak electric current flows through the electromagnet. The electromagnet then pulls the piece of soft iron, thereby closing a circuit in which a much stronger current can flow.

Fig. 18.45
A relay

You most likely know one of the applications of relays. When the key of the ignition of a car is turned all the way, the starter motor begins to run. (The starter is an electric motor that actuates the internal combustion engine of the car. It receives energy from the car battery). A strong electric current of about 100 A flows through the starter motor. A large and robust switch is necessary for this. Such a switch would be too big to fit into the ignition lock. For this reason the starter current is turned on and off through a relay. The weak control circuit of the relay can be turned on and off with a small switch in the ignition.

Exercises

1. What appliances or devices are electromagnets used in? Name some that do not appear in the text.

2. Invent an electric bell that functions with alternating currents only.

3. How does the ammeter in Fig. 18.44 react to an alternating current? Invent an alternating current ammeter.

18.11 Electric motors

We want to build an electric motor ourselves. We will start with a primitive version. Our motor is similar to the device for measuring currents in Fig. 18.44.

Fig. 18.46 shows it from above. There are electromagnets on the left and on the right. In between them there is a permanent, rotary magnet. We turn on the electric current. Poles form on the face surfaces of the electromagnets. The permanent magnet in the middle now turns so that the unlike poles move together as closely as possible, Fig. 18.46a. As soon as the permanent magnet reaches a position parallel to the electromagnets, we interchange the connections with the battery terminals. This causes the poles of the electromagnets to be reversed. The fields now repel the adjacent poles. The bar magnet continues to rotate, Fig. 18.46b. We reverse poles again after a half turn, and so on. In this way, the fields keep the permanent magnet in constant rotation.

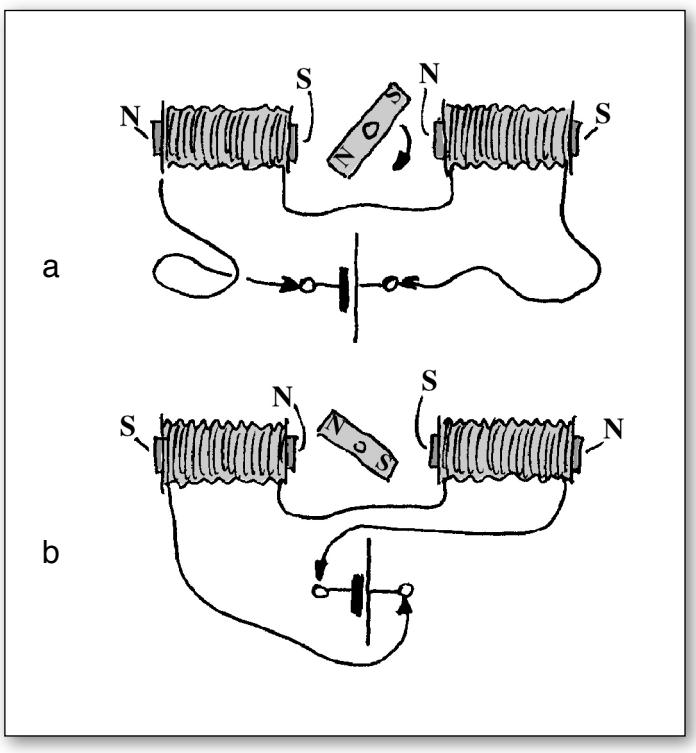

Fig. 18.46
Principle of operation of an electric motor

You will notice that it is difficult to reverse the polarity of the electromagnets at exactly the right moment, and a hand controlled electric motor is pretty useless anyway. We need a motor that can control itself, a motor that automatically reverses the direction of the electric current in its solenoids after every half rotation.

It is pretty simple to build an automatic control. We only need to install a switch on the motor's axle that is activated by its rotation.

It is particularly convenient to realize this kind of switching if the roles of the permanent magnet and the electromagnet are reversed. The electromagnet is made to rotate and the permanent magnet is fixed, Fig. 18.47. The electromagnet is then called the rotor of the motor. Electrical inflow and outflow both occur through two sliding contacts and a sliding ring that is divided into two halves insulated from each other. The electromagnet is connected to these two halves.

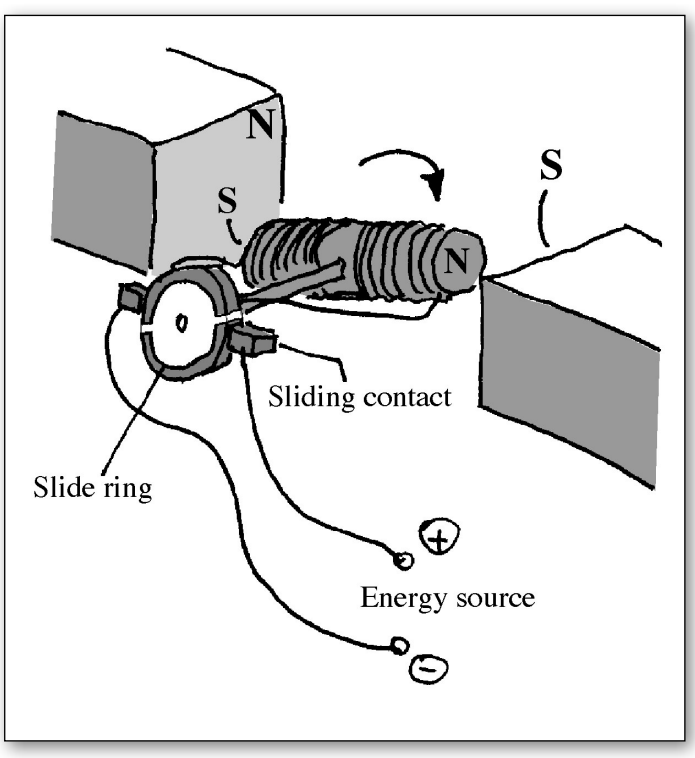

Fig. 18.47
An electric motor

Every time the rotor makes a half rotation, the halves of the sliding ring change from one sliding contact to the other. In the process, the electromagnet's terminal that initially is at a high potential, now is at a low potential, and the one that had a low potential now has a high potential. In this way, the current in the electromagnet always reverses at exactly the right moment.

Many electric motors function by this principle. There are, however, a lot of other tricks that can be used when building electric motors, although they all have one thing in common: It is always a magnetic field that attracts or repels the rotor.

Exercises

1. The "motor" in Fig. 18.46 can be used as an alternating current motor. It is unnecessary to reverse its polarity by hand. This type of motor is called a synchronous motor. What problems would occur with it?

2. Design an electric motor where both the fixed and the rotating magnets are electromagnets.

18.12 The Earth's magnetic field

We already saw that if a bar magnet is hung from a vertical axis so that is can rotate easily, it will position itself approximately north to south. One pole will point north, the other south, Fig. 18.48. The end pointing north is called the north pole and the end pointing south is called the south pole.

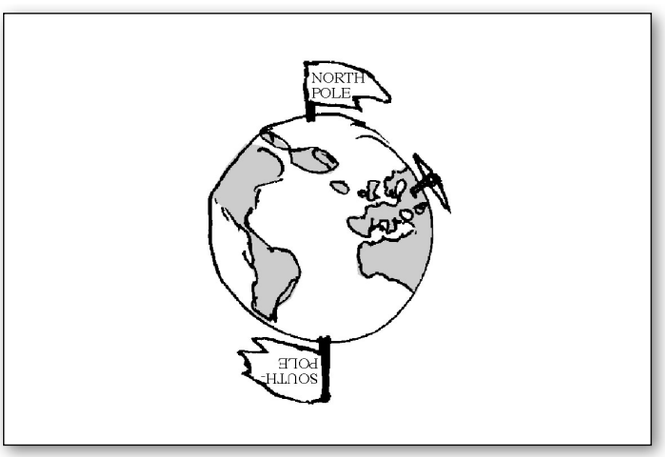

Fig. 18.48
A rotary magnet will adjust to a north-south direction.

Compasses are based on this principle. The compass needle is simply a light rotary permanent magnet.

Actually, compass needles don't show exact north and south. Moreover, the deviations from this north-south direction can vary at different places on Earth. They even change slowly with time.

In any case, we can conclude that the Earth is surrounded by a magnetic field. It has been discovered that this magnetic field reaches into the Earth as well.

Where does this field come from? Two causes come to mind easily. Either the Earth itself is a huge permanent magnet, or there are electric currents flowing through it. In earlier times, people believed the first of these hypotheses to be true. It was thought that the Earth actually was a huge permanent magnet. In that case, the north pole charge would have been at the geographic south pole of the Earth and vice versa because a magnet's north pole charge is attracted to the north.

In the last century, however, it was realized that this assumption was wrong. Inside the Earth it is so hot that every material loses its magnetism. Therefore, only electric currents can be the cause of the Earth's magnetic field.

Exercises

1. Why does a compass show wrong directions when it is near pieces of iron?

2. Two compass needles are put very close together. What directions do they show?

18.13 Induction

A voltmeter is connected to a solenoid. If a permanent magnet is moved into the solenoid, Fig. 18.49, the voltmeter's needle moves sharply, but only as long as the magnet moves. If the magnet is removed from the solenoid, the measuring device's needle once again moves sharply, but this time it moves in the opposite direction.

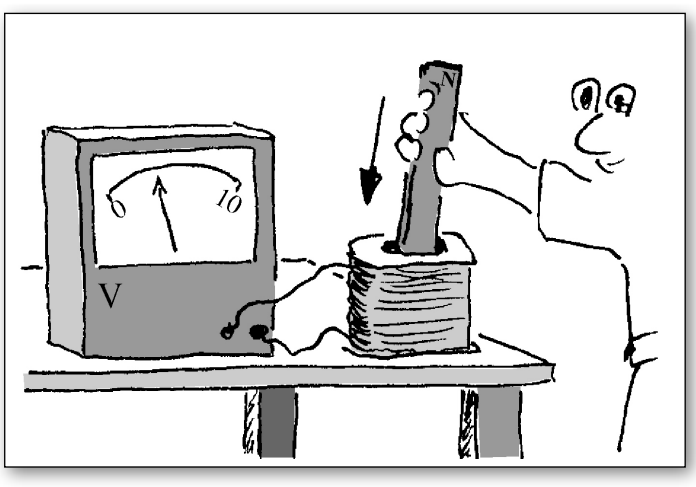

Fig. 18.49
The voltmeter shows a reading as long as the permanent magnet is moving.

The needle's movement depends upon whether the north or the south pole is put into the solenoid.

The magnetic field inside the solenoid changes with the magnet's movement. This change in the magnetic field is responsible for the voltage between the terminals of the solenoid.

We now ask what happens if the solenoid is short-circuited during the experiment. We short-circuit it, but attach an ammeter to the wire that we use to bridge the two terminals of the solenoid. The result is that the needle of the ammeter moves sharply when the magnet is pushed in and again when it is slid out, as you certainly would have expected.

These processes are called *induction*. One says that during the motion of the magnet a voltage or an electric current is *induced*.

The more strongly the field in the solenoid changes, the greater the induced voltage is. We now wish to try creating the highest possible induced voltage. To do this, we need only make sure that the field in the solenoid changes as fast and as much as possible.

First we notice that the voltage is higher the faster the permanent magnet moves. Eventually however, the effect doesn't increase anymore upon accelerating the movement. This is because the measuring device cannot follow the motion anymore. It is too slow for this. If a so-called oscilloscope is used instead, it is possible to see that the fast movement causes the voltage to rise still more.

Next we make the change of the field even stronger by inserting two magnets into the solenoid so that the like poles are next to each other, Fig. 18.50.

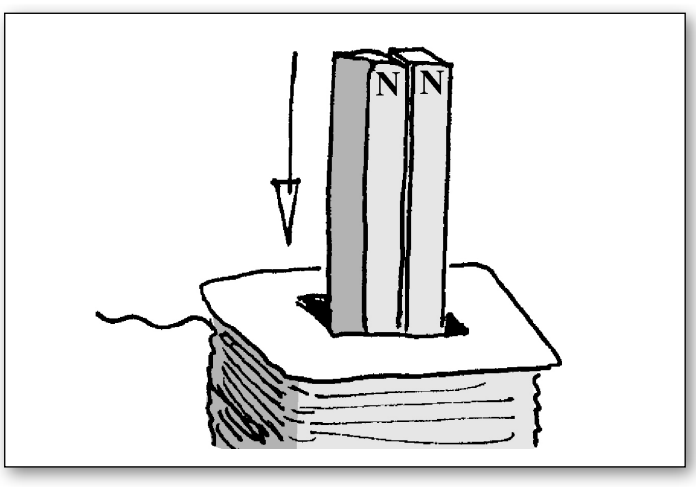

Fig. 18.50
If the magnetic field is denser, a greater voltage is induced.

A third method to strengthen the induction effect is to use a solenoid with more coils.

There is another completely different way of changing the field in the solenoid, a way in which nothing moves. An electromagnet is put next to the solenoid so that its field reaches into the solenoid, Fig. 18.51. If the electromagnet is turned on or off, the magnetic field in the solenoid is changed and a voltage is induced.

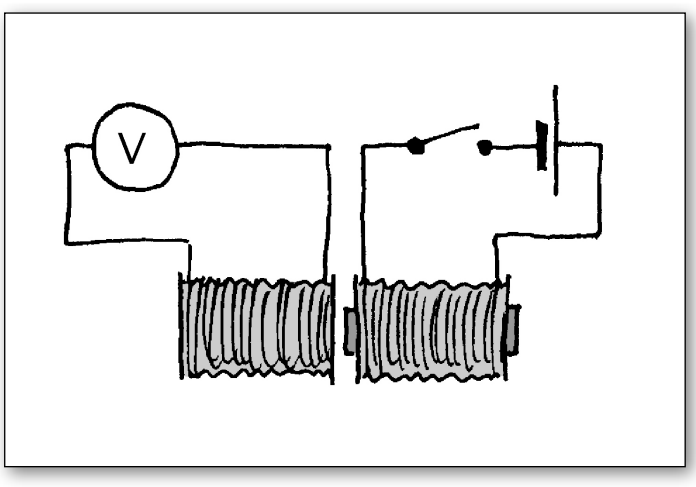

Fig. 18.51
The density of the magnetic field in the coil changes when the electromagnet is turned on or off, thereby inducing a voltage.

A change in the magnetic field of a solenoid produces a voltage between the terminals of the solenoid. In a closed circuit, an electric current flows. This process is called induction.

Finally, we will do another variation of the induction experiment. We insert a soft iron core into the solenoid and lengthen the ends of this core so that it makes a "U" shape. No magnetic field can penetrate into the solenoid now. We move a permanent magnet near the ends of the soft iron core, Fig. 18.52, until the poles of the magnet touch the ends of the "U" shaped iron core. We observe that the voltmeter's needle moves sharply. How is this possible? Since the solenoid is filled with soft iron, there is no field and no change of a field.

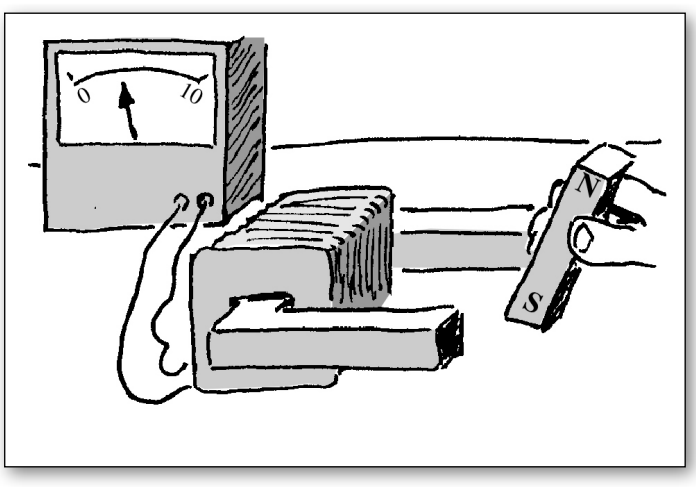

Fig. 18.52
Changing the magnetization inside the solenoid also causes an induced voltage.

Nevertheless, something happened within the solenoid. The iron in the solenoid has been magnetized, and its magnetization has changed.

We now turn the permanent magnet so that its north pole is where its south pole was, and vice versa. This process also induces a voltage in the solenoid. The result of this experiment is:

If the magnetization of the material in a solenoid changes, a voltage (a current) is induced.

The functionality of many important devices is based upon induction. In the following, we will get to know some of these devices.

Exercises

1. A voltage should be induced in a solenoid with the help of a permanent magnet. How can the voltage be made as high as possible? Name three ways of doing this.

2. A solenoid is held so that its axis is vertical so that we can let an object fall through it. An oscilloscope is connected to the solenoid. A bar magnet is let fall lengthwise through the solenoid. What does the oscilloscope show?

18.14 Generators

A generator does exactly the opposite of what an electric motor does. An electric motor receives energy with the carrier electricity, and emits energy with the carrier angular momentum, Fig. 18.53a, while a generator receives energy with the carrier angular momentum and emits it with the carrier electricity, Fig. 18.53b.

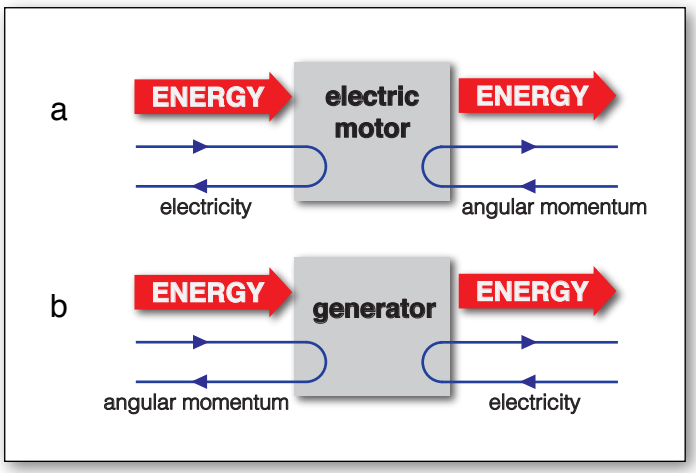

Fig. 18.53
Flow diagram of an electric motor and a generator.

A generator's construction is basically not different from that of an electric motor. We could use the motor sketched in Fig. 18.47 as a generator. We would need only to replace the electric energy source with an energy receiver, maybe a light bulb. When the axle is rotated very quickly, the lamp starts to burn. It is fairly simple to explain how a generator works. The iron core of the rotary solenoid is remagnetized by the fixed permanent magnet twice per rotation. Each time a voltage is induced between the ends of the solenoid. The sign of the voltage changes twice per rotation. The sliding contact and the two-part sliding ring ensure that there is always a voltage of the same sign across the generator's terminals. (This voltage is not constant in time, it is not a true direct voltage.)

It is even simpler to build an alternating voltage generator. Do you know how to do this?

Generators are some of the most important machines in any power plant.

Generators can sometimes have other names. For example, a bicycle's or a cars generator is called a dynamo.

18.15 Transformers

Many electric devices such as transistor radios, cassette players and toy motors use much smaller voltages than the 117 V in wall outlets. If these devices are to be connected to this high voltage, it is necessary to reduce the 117 V to a smaller value. It needs to be "transformed" to a smaller value. There is something that can be connected between the device and the wall outlet that does this. It is called a *transformer*.

A transformer is made up of an iron core with two solenoids or coils, Fig. 18.54. One of these coils, the so-called primary coil, is connected to the wall outlet. The other, secondary, coil is connected to the energy receiver which is the device being used.

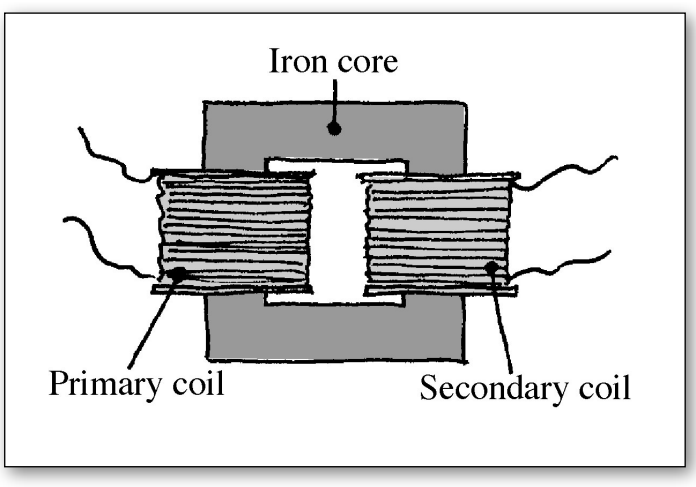

Fig. 18.54
How a transformer is constructed

We wish to familiarize ourselves with how a transformer works. We connect one coil of a transformer to a light bulb and the other through a switch to a battery, Fig. 18.55. When the switch is closed, the light bulb lights up shortly. When the switch is opened, the light bulb again lights up for a moment. There is an easy explanation for this. The magnetization changes everywhere in the iron core when the switch is turned on or off. Because it changes in the secondary solenoid as well, a voltage is induced there.

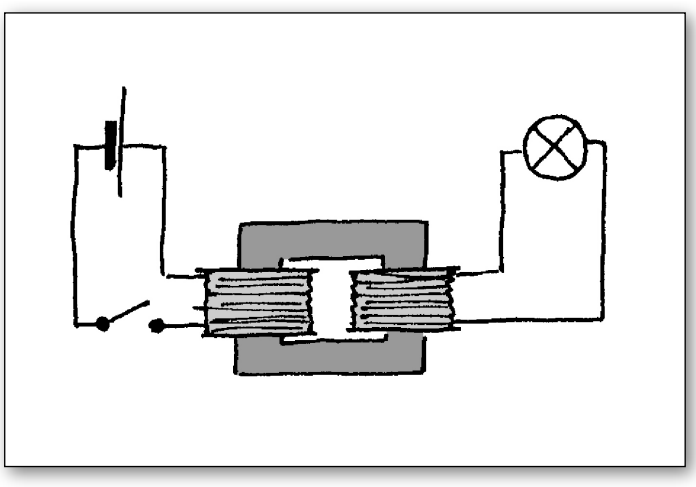

Fig. 18.55
The lamp burns shortly when the switch is turned on or off.

In order to keep a lamp burning constantly, the electric current in the primary coil should be turned on and off in quick succession. Instead, the primary coil can be simply connected to an alternating voltage source. The voltage induced in the secondary coil is then, of course, an alternating voltage. We now see that a transformer works only with an alternating voltage.

How can a voltage be brought up or down with a transformer? The value of the induced voltage depends upon the number of turns of the wire in both coils. We want to investigate how.

We construct transformers out of coils made up of different numbers of turns of the wire. We notice that if the number of turns in both primary and secondary coils is the same, then primary and secondary voltages are the same. If the secondary coil has twice the number of turns of the primary coil, the secondary voltage is twice that of the primary voltage. In general:

$$\frac{U_1}{U_2} = \frac{n_1}{n_2}$$

where U_1 and U_2 are the voltages across the primary and secondary coils, and n_1 and n_2 are the corresponding turns.

While the transformer's electric current flows in two separate circuits, Fig. 18.56a, the energy flows from the primary side to the secondary side of the transformer, Fig. 18.56b.

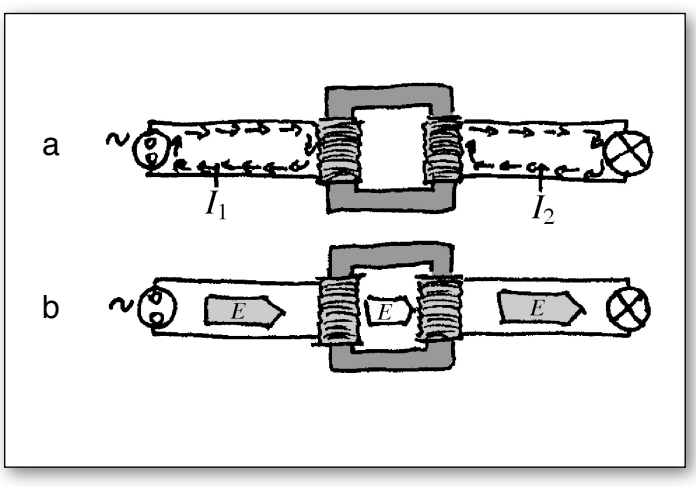

Fig. 18.56
Paths of the electric current (a) and the energy current (b) in a transformer

The energy current P_1 flowing into the transformer is, except for small losses, the same as the energy current P_2 leaving the transformer. Therefore

$P_1 = P_2$.

And because $P = U \cdot I$,

$U_1 \cdot I_1 = U_2 \cdot I_2$ must be true as well.

U_1, U_2, I_1 and I_2 are the voltages and currents of the primary and secondary circuits, respectively.

We conclude from the last equation that a reduction of the voltage in a transformer is associated with an increase of the electric current by the same factor. Likewise, an increase of the voltage corresponds, by the same factor, to a reduction of the electric current.

Exercises

1. The coils of a transformer have 1000 and 5000 turns, respectively. There is an alternating voltage of 117 V being used. What voltages can this transformer create?

2. The primary solenoid of a transformer is connected to a wall outlet. A voltage of 12 V is measured at the secondary solenoid. What can be said about the number of turns of the transformer's coils? An electric current of 2 A is flowing in the secondary electric circuit. What is the current in the primary circuit?

3. A transformer has a primary coil with 1000 turns and a secondary coil with 10,000 turns. The primary coil is connected to a wall outlet. A primary current of 100 mA flows. What are the secondary voltage and the secondary currents?

4. An energy current of 100 kW flows through the transformer in Fig. 18.57. What are the requirements for the inlet cable (left)? What are the requirements for the outlet cable (right)? Why is it preferable to transport energy carried by electricity using high voltages?

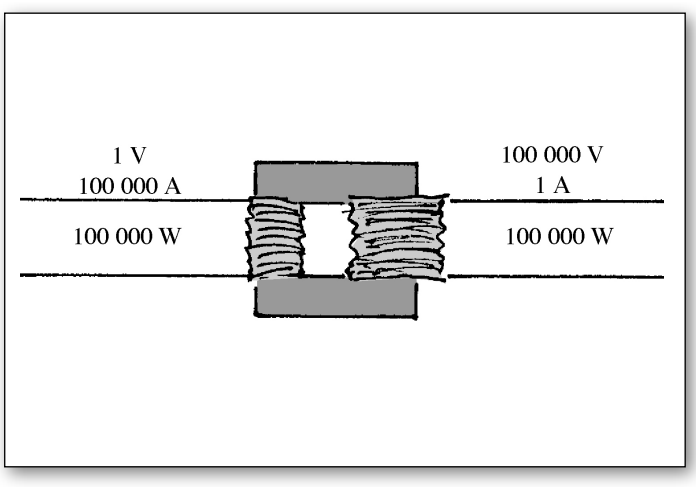

Fig. 18.57
For Exercise **4**

18.16 The magnetic field of induced currents

Turn back to section 18.10 where we saw that an electromagnet
– can attract or repel a permanent magnet, Fig. 18.44;
– can attract or repel another electromagnet, Fig. 18.46;
– always attracts a piece of soft iron, Fig. 18.45.

We will now do some experiments that are very similar to those described in section 18.10 (see Fig. 18.58). We suspend a conducting ring having a slit next to an electromagnet. The ring is allowed to move easily. It can be made of aluminum or copper because the material cannot be magnetic. We now turn on the electromagnet. Nothing happens – as you probably expected. A piece of matter without a current flowing in it is no magnet.

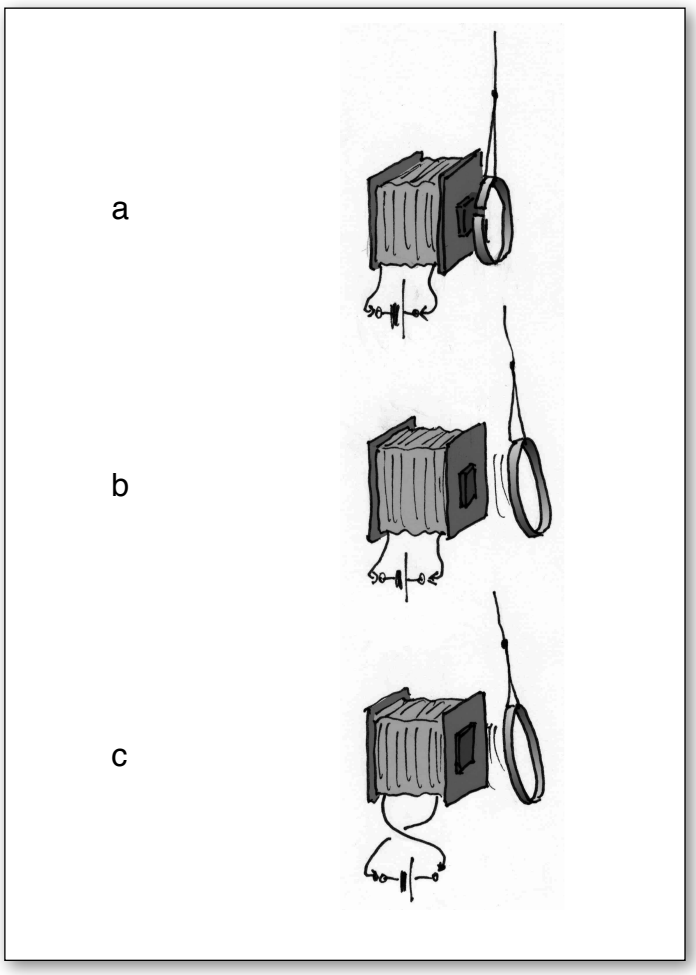

Fig. 18.58
(a) A ring with a slit will be neither attracted nor repelled.
(b, c) A closed ring is always repelled when the electromagnet is turned on no matter in which direction the current flows.

We now make a little change: We take a ring without a slit and again turn on the electromagnet. This results in the ring being repelled. How does this happen?

This can only be explained by the ring having an electric current flowing through it. How does this current appear in the ring? Through induction, of course. When the electromagnet is turned on, the magnetic field in the ring is changed so that an electric current is induced. This current creates a magnetic field. Both fields together – that of the electromagnet and that of the ring – cause the repulsion. This effect disappears immediately, however, because the induced field exists for only a short time during which the electromagnet is turned on.

We will repeat the experiment but this time we exchange the terminals of the electromagnet. This time, the current flows in the opposite direction when it is turned on. Our observation: The ring is repelled once again. Surprise! Or maybe not? Actually, this could be expected. We have reversed the polarity of the electromagnet and reversed the direction of the induced current. Therefore we have inverted both of the “magnets” (electromagnet and ring), and, naturally, repulsion did not change into attraction.

We can now formulate a rule:

> The direction of the induced current in a ring is such that repulsion results between the ring and the magnet causing the induction.

18.17 Superconductors

There are materials that lose their electric resistance when they are cooled down below a certain temperature. They are called *superconductors*. The transition temperature from normal to superconducting states is relatively high for some materials, around –180 °C. These materials can be brought rather easily to a superconducting state by cooling them with liquid nitrogen.

Superconductors are not only interesting because they have no electric resistance. They also have surprising magnetic characteristics.

We construct an arrangement of permanent magnets whose field points upward and has a dent in the middle, Fig. 18.59. We put a small piece of superconducting material close to the magnets, and then let go of it. The superconductor does not fall.

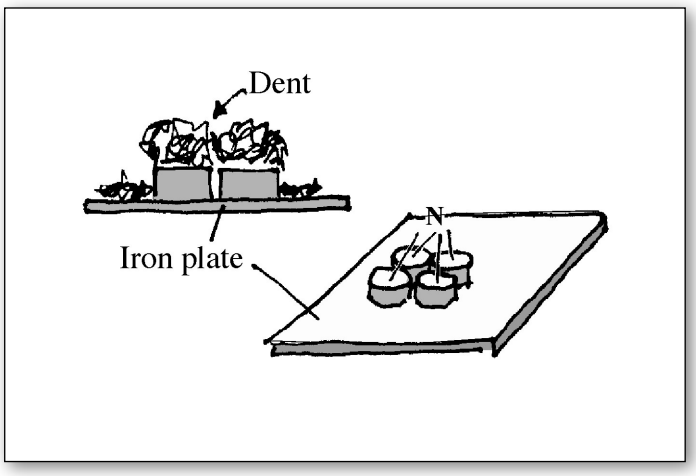

Fig. 18.59
The magnetic field above has a dent.

It floats above the magnets Fig. 18.60. We can turn it or push it a bit to the side, but it remains floating (naturally only until it has warmed up and returned to its normal state).

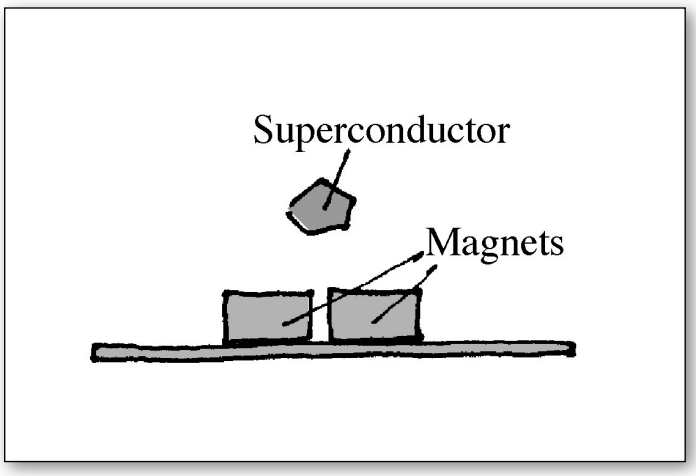

Fig. 18.60
The superconductor is kept floating by the magnetic field.

The superconductor is obviously being repelled by the magnets. It behaves the opposite of a piece of soft iron which would always be attracted to the magnets. How can this be explained? We remember the result of the last section. We observed there that the ring is always repelled by the electromagnet.

The explanation for the superconductor being repelled is the same. As soon as a super conductor comes near a magnet, currents begin to flow in it and these are oriented so that repulsion occurs. Induced currents in the ring rapidly stopped flowing because they were counteracted by the ring's resistance. In the case of the superconductor, the activated currents flow on as long as it is suspended over the magnets. There is no resistance to these currents.

A more extended investigation that we cannot perform here also shows,

– that the currents only flow very close to the surface of a superconductor;

– and that the magnetic field does not penetrate into a superconductor.

Superconductors have characteristics similar to those of soft magnetic materials. They do not let the magnetic field penetrate either. They do this with another "trick" however. They activate electric currents instead of creating magnetic poles.

19

Electrostatics

19.1 Charge and charge carriers

When electricity moves from one side to the other in a wire, we talk about electric currents. Until now, we have dealt with the effects of electric currents and with the relation between electric currents and other quantities. However, we have never asked about the effects and characteristics of electricity itself. Ideally, the easiest way to investigate electricity would be if the electricity were motionless, meaning when no electric current is flowing.

Admittedly, one notices almost nothing of the electricity in a copper wire not connected to a circuit. Why is this? A possible answer would be that electricity at rest has no characteristics that make it noticeable. This is wrong. Electricity makes itself clearly noticeable even when it appears in very small amounts. The field of *electrostatics* describes this. The fact that it is unnoticeable in a copper wire is because of a characteristic that makes electricity different from many other quantities. It can assume both positive and negative values.

All material substances contain electricity, but they almost always contain equal amounts of positive and negative electricity, so the total amount is zero. 1 g of copper contains 44,032 C of positive electricity. It also contains the same amount of negative electricity, bringing the total amount to 0 C.

(By comparison, mass, meaning amounts that are measured in kg, can only have a positive value.)

Electricity can take positive and negative values.

What sense does it make to say that a body with a total amount of electricity of 0 C, actually has a certain amount of positive and an equal amount of negative electricity? 0 C means nothing more than that it has *no* electricity in it. That it does make sense to say that copper (or any other substance) contains positive as well as negative electricity, can be recognized when the microscopic structure of matter is considered.

All substances are composed of atoms and groups of atoms called molecules. Every atom is made up of protons and neutrons combined in the nucleus, and electron shells. Two of these components carry electricity. The proton carries positive electricity, in fact

$Q_{\text{Proton}} = 1.602 \cdot 10^{-19}\,\text{C}$.

The electron carries negative electricity, namely

$Q_{\text{Electron}} = -1.602 \cdot 10^{-19}\,\text{C}$.

Neutrons carry no electricity, therefore

$Q_{\text{Neutron}} = 0\ \text{C}$.

Because an atom has equal amounts of protons and electrons, the total amount of electricity carried by the atom is 0 C.

It can happen that an atom has one or more electrons too few or too many. This entity is called an *ion*. The amount of electricity in an ion is not zero.

We have now learned about another important characteristic of electricity. It always sits upon some particle. Along with protons and electrons, there are numerous other charged particles such as positrons, muons, antiprotons, and more. Under normal circumstances, these usually do not exist, but they can be artificially produced. However, they have a very short life span.

Particles that have a charge on them are said to be *electrically charged*. For this reason, it has become common to call electricity *electric charge*, and electrically charged particles (electrons, protons, ions, etc.), *charge carriers*.

Electric charge (electricity) is always on particles, the charge carriers.

19.2 Charge currents and charge carrier currents

Now we can understand what distinguishes an electric conductor from a non-conductor. Conductors are substances that contain *mobile* charge carriers, while in non-conductors, the charge carriers are *immobile*. Which particles make up the mobile carriers, varies from case to case. In some conductors, only positive charge carriers are mobile. In others, it is only the negative ones. In yet others, both positive and negative charge carriers are mobile.

In metals, the mobile charge carriers are electrons. However, not all the electrons of a metal atom are mobile. Only one per atom is. There are no mobile electrons in acids, bases, or saline solutions. In these cases, conductivity comes from the mobility of ions. Because there are positive as well as negative ions, we have charge carriers for charges of both signs.

If an electric current flows in a circuit, the mobile charge carriers push past the other, oppositely charged, ones. This leaves the electric circuit neutral everywhere: Conductors, energy sources, and energy receivers remain uncharged.

We see that an electric current can be produced in different ways. In all three parts of Fig. 19.1, we have an electric current of 2 A flowing from left to right. In part (a) of the figure, the current is created by positively charged carriers moving from left to right. In (b), negative charge carriers flow from right to left. In part (c), the positive charge carriers move right while at the same time, negative charge carriers move to the left. Both types of charge carriers contribute to the total current.

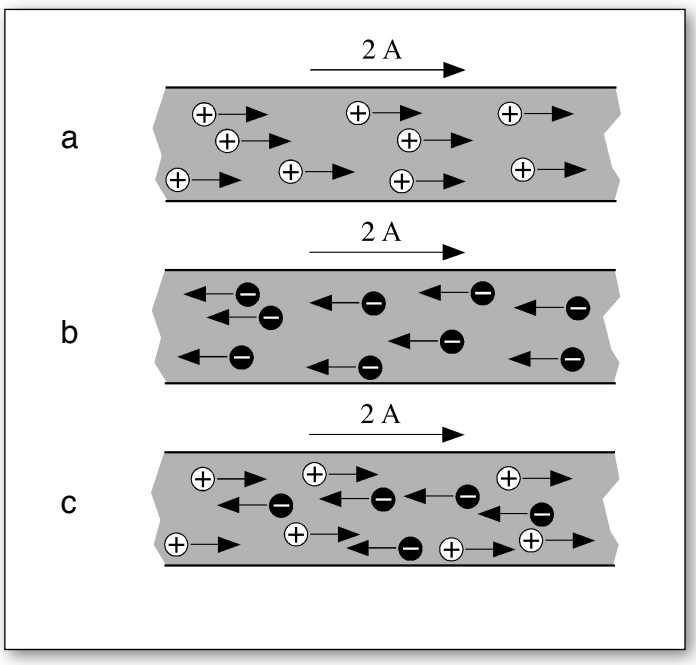

Fig. 19.1
An electric current results from charge carriers that move (a) to the right, (b) to the left and (c) in both directions.

You may be surprised to find out just how slowly the charge carriers in a conductor move. If an electric current of 1 A flows in a copper wire with a cross section of 1 mm^2 , the mobile charge carriers (mobile electrons) move at a speed of 0.07 mm/s.

It is possible to make the movement of charge carriers visible. An electric current is let flow through a saline solution, Fig. 19.2. A solution of potassium nitrate is in a shallow groove on the left, and on the right, a solution of potassium permanganate. The solution on the left is clear and transparent, the one on the right is violet. The coloration of the potassium permanganate solution is due to the negative permanganate ions. Now, if an electric current flows, all the ions move, including the permanganate ions. The movement of the permanganate ions results in a shifting of the boundary between the clear and the violet areas, and this is easy to observe. If the electric current flows from left to right (Fig. 19.2), the violet ions move to the left. If the electric current flows to the left, they move to the right.

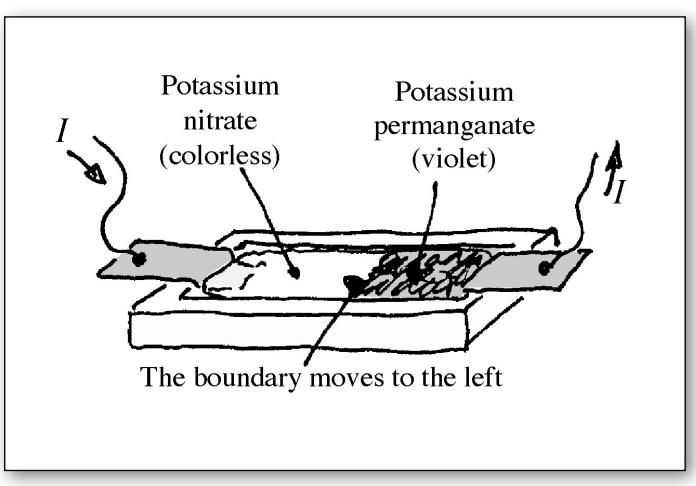

Fig. 19.2
If an electric current flows through the solution, the boundary between the violet potassium permanganate and the colorless potassium nitrate solution shifts.

Exercises

1. Positive ions flow from left to right in a saline solution with two electrodes immersed in it. They transport 0.5 Coulomb per second. At the same time, negative ions flow from right to left. They transport –0.3 Coulomb per second from right to left. What direction does the electric current flow in? How strong is it?

2. An electric current of 2 A flows through a copper wire. How many electrons per second move through a cross section of the wire?

19.3 Accumulation of electric charge

Our first goal was to learn something about the characteristics of electricity. We found out why an electric circuit is electrically neutral everywhere, why it is usually impossible to notice electric charge. We now wish to upset the charge balance of an electric conductor. We will try to accumulate charge on it so that its total amount of charge is no longer zero. We will see that it is rather difficult to do this.

In order to better understand the problem we are dealing with, we consider Fig. 19.3. The container on the left is filled with air under normal pressure. We wish to increase the amount of air in this container, so we simply pump in air from outside. In the process, the pressure increases. The container on the right in Fig. 19.3 is filled with water, and we want to increase the amount of water in it. This is far more difficult than it is with air. Even a pump that can produce extremely high pressures can only increase the amount of water very slightly. This is because water cannot be compressed as easily as air.

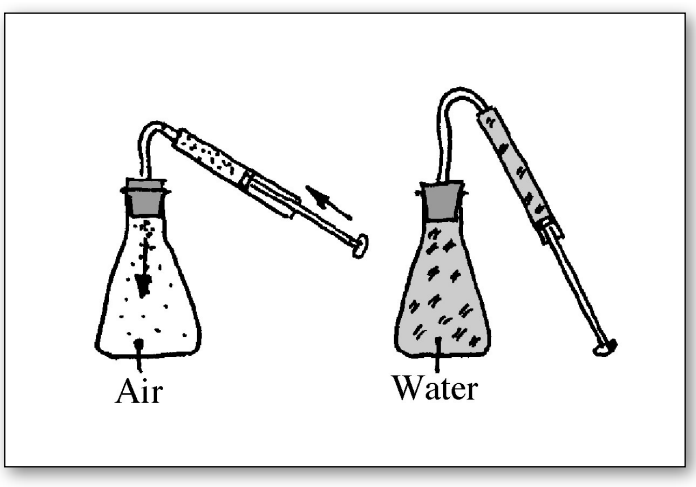

Fig. 19.3
The amount of air in the container on the left can be easily altered. The amount of water in the container on the right is difficult to change.

Electricity behaves similarly to water in this respect. It is very difficult to change the amount of electricity in an object from the normal value (0 Coulomb).

How can we accomplish an accumulation of electricity in an object at all? With an “electricity pump”, of course, i.e., with a battery or a power supply. Fig. 19.4 shows an experiment that fails. The positive terminal of a battery is connected to a wire and the negative terminal is grounded. The battery should take electricity out of the Earth and push it into the wire. The wire should become electrically charged and stay that way when it is disconnected from the battery. If it is touched by a connection to a lamp whose other connection is grounded, the lamp should burn for a moment because the accumulated electricity should flow over the lamp and back into the Earth. However, the lamp does not burn. Why not? Because the amount of electricity that has been pumped onto the wire was much too small.

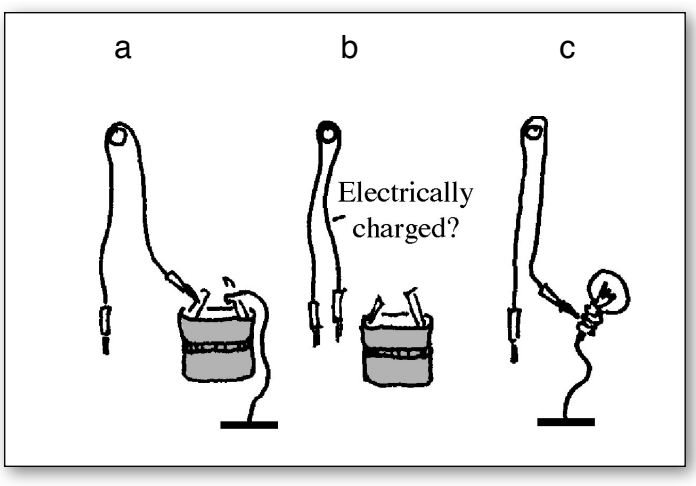

Fig. 19.4
(a) The battery pumps electricity out of the Earth into the wire.
(b) The wire is electrically charged.
(c) The lamp does not light up because the charge on the wire is much too small.

In order to demonstrate an accumulation of charge on a wire, we must improve our experiment in three ways:

(1) We use an “electricity pump that pumps much more strongly”. This would be a power supply that creates a much higher voltage. A normal high voltage power supply (with a transformer) or a Van de Graaff generator could be used, Fig. 19.5. A Van de Graaff generator creates voltages of up to 50 kV.

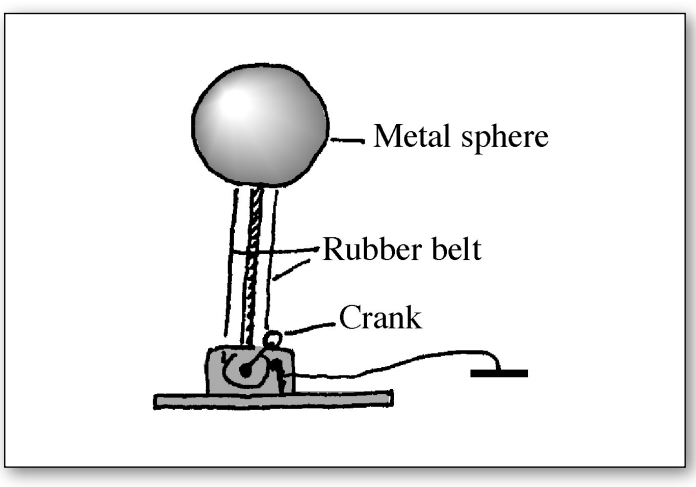

Fig. 19.5
Van de Graaff generator

(2) To show the charge of the wire, we use a device that is more sensitive and reacts to smaller amounts of charge than a light bulb. It is called a glow lamp. A glow lamp has an additional advantage for our experiment. It also shows the direction of the current flowing through it. It always glows on the side where the lower potential lies.

(3) We suspend the wire over insulators. The normal plastic insulation of a wire is so bad that electricity accumulated on it could flow through this insulation and the table into the ground.

After we have taken these measures, our charge accumulation experiment is successful. Depending upon which of the terminals of the high voltage power supply is grounded, the amount of electricity in the wire increases or decreases. If the negative terminal of the power supply is grounded, Fig. 19.6, the wire becomes positively charged. The fact that the mobile charge carriers of the wire are electrons means that there are some electrons missing in the wire. It has fewer electrons than in its uncharged state. If the positive terminal is grounded, and the negative terminal is connected to the wire, the wire becomes negative. It has an excess of electrons.

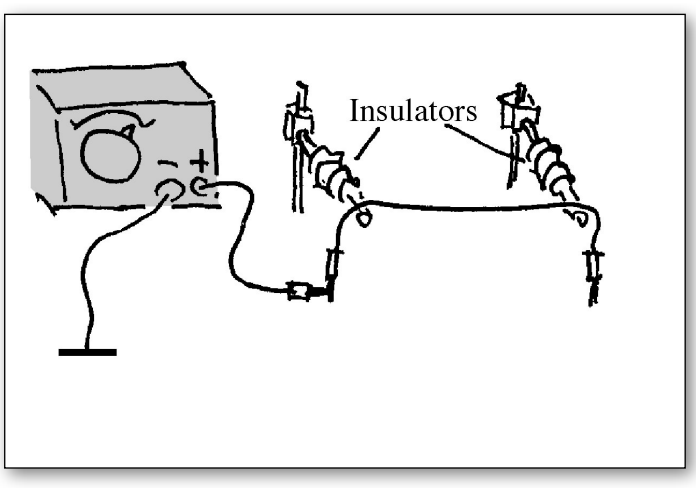

Fig. 19.6
The wire has positive charge, there is a lack of electrons.

The amount of charge accumulated on the wire is larger the higher the electric potential it is brought to. A high potential is related to a (relatively) large quantity of positive charge. A high negative potential leads to a (relatively) large quantity of negative charge.

> The higher the electric potential of a body, the more electric charge it contains.

The reverse is also true:

> The larger the amount of electric charge on a body, the higher the electric potential.

19.4 Electric fields

We have now managed to accumulate charge and to demonstrate the accumulation. However, we have not noticed any special characteristics of electric charge. In order to investigate the characteristics of electricity, we perform the experiment sketched in Fig. 19.7. Two metal spheres A and B are connected to a high voltage power supply. Sphere B is very light. It is suspended from a thin wire so that it can move. If the power supply is now turned on so that one sphere becomes positively and the other negatively charged, B is attracted by A. If we reverse the signs of the charge, nothing changes: B is again attracted by A.

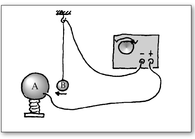

Fig. 19.7
The electric field pulls sphere B toward sphere A.

Now we connect the spheres to the power supply so that their charges have the same sign, Fig. 19.8. B is now pushed away by A. It makes no difference whether they are both positively or both negatively charged.

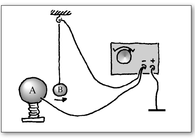

Fig. 19.8
Sphere A pushes sphere B away by means of the electric field.

The fact that one sphere is attracted to the other allows us to conclude that here is a connection between them.

This connection is called an *electric field.*

> Electrically charged objects are surrounded by an electric field. If the charges of two objects have the same sign, the field pushes the objects away from each other. If they have different signs, the field pulls them toward each other.

We will do a last experiment that is simpler than the ones we have just performed, Fig. 19.9. Only the fixed sphere A is electrically charged, and sphere B is insulated and suspended. Surprisingly, B is attracted to A. It does not matter whether A is positively or negatively charged. How can this be explained? The movable sphere is not connected to the power supply so it cannot have a field surrounding it.

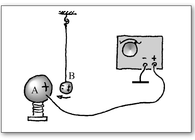

Fig. 19.9
The mobile charge carriers on B are displaced by the electric field. Electrically charged areas form on the surface of B.

We can find the explanation for this if we recall a similar phenomenon of magnetism. A piece of soft iron, meaning an object that is non-magnetic at first, will be attracted to a magnetic pole. This can be either the north pole or the south pole. This is because soft iron forms poles when it comes near a magnetic pole. The side near the north pole forms a south pole, and a north pole forms on its other side, Fig. 19.10.

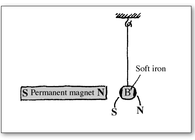

Fig. 19.10
B is magnetized by the magnetic field. Magnetic poles form on the surface of B.

This is quite similar to our last experiment. The electric field pulls on the charge carriers of B and shifts them slightly so that one side of B becomes positively and the opposite side, negatively charged. The total charge remains zero. If A is positively charged, the side of B facing A becomes negatively charged. The side of B facing away from A becomes positively charged. Because the negative side of B is closer to A than its positive side, sphere B is attracted to sphere A.

If A is negative, the charges on B are reversed and once again, A's charge and the charge on the side of B facing it, are opposite. This again causes B to be pulled toward A.

This process of charge transfer under the influence of another body is called *electrostatic induction.*

Earlier we had only a glow lamp for proving that an object is charged. Another device for demonstrating charge is the *electroscope.* It works as follows.

There is a vertical rod inside a ring, Fig. 19.11. This rod is electrically insulated from the ring. There is a rotary rod attached to the first rod. It is very light and is connected conductively to the first one. Both rods are connected conductively to the electroscope's upper terminal (see the figure). The ring is grounded.

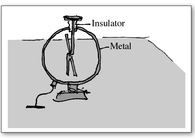

Fig. 19.11
An electroscope. The movable rod carries charge with the same sign as the fixed one.

We shall use the electroscope to demonstrate that there is charge on the sphere. We touch the upper terminal of the electroscope with the sphere. Some electric charge flows from the sphere to the two rods. They now carry like charges, and the rotary rod is repelled by the fixed rod. The more charge there is on the electroscope, the more the rotary rod moves away from the fixed one.

We will use the electroscope to do a simple experiment to show, once again, the phenomenon of electrostatic induction, Fig. 19.12.

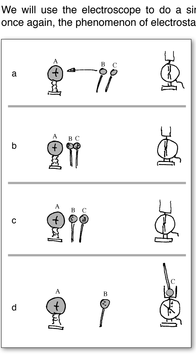

Fig. 19.12
(a) The neutral spheres B and C are put near A.
(b) The charges on B and C split by electrostatic induction.
(c) The contact between B and C is interrupted.
(d) The charge on B and C can be demonstrated by an electroscope.

The large sphere (A) is positively charged. We put two small neutral spheres B and C inside the field of the large one, Fig. 19.12a. B and C touch each other but they do not touch A, Fig. 19.12b. The field of sphere A causes the charges on B and C to divide: we have electrostatic induction. Negative charge accumulates on the left, on sphere B. The positive charge concentrates on the right, on sphere C. We separate B and C from each other while they are near A, Fig. 19.12c. We then move them away from the large sphere's influence, Fig. 19.12d. The charges on B and C should actually neutralize each other, but they do not because the connection has been severed.

We use an electroscope to show the charges on B and C. We touch the electroscope with one of the spheres, B for example. Negative charge flows from B to the electroscope and it shows a reading. We then touch sphere C to the electroscope, and positive electricity flows to the electroscope neutralizing the negative. The reading goes back to zero.

Exercises

1. The positive and negative charge carriers on sphere B in Fig.19.9, are separated by electrostatic induction. The field attracts the sphere to A. However, as soon as it touches A, it is repelled. How can this be explained?

2. How could one demonstrate that the entity in the vicinity of electrically charged objects is not a magnetic field?

3. A light metal sphere A is suspended between two fixed spheres B and C, Fig. 19.13. Sphere A is brought into contact with sphere B for a short moment. It is then let go. What happens?

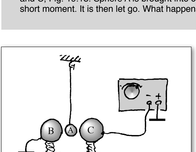

Fig. 19.13
For exercise **3**

19.5 Capacitors

We have charged a wire and a metal sphere. We could only accumulate very little charge despite the high potential. We have not yet investigated the relation between amount of charge and size and form of a charged body. We will do it now. It is our goal here to find out how to store as much charge as possible.

Our first conclusion is that electricity lies only upon the outer surface of a charged body. We recognize this because at the same potential, the amount of charge on a solid sphere is the same as on an hollow one of the same size, Fig. 19.14. However, a large sphere has more charge on it than a small one. An object with a large outer surface is needed when a lot of charge is to be stored.

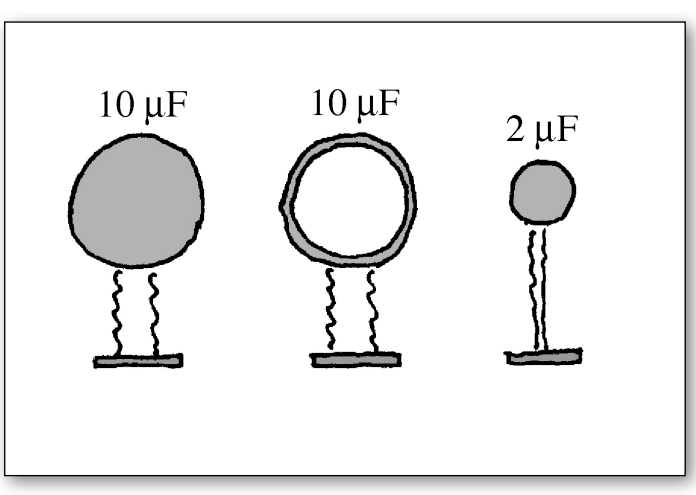

Fig. 19.14
At the same potential, the amount of electricity on a hollow sphere is the same as that on a massive one, but a smaller sphere has less electricity on it than a larger one.

There is a much more effective method for increasing the amount of accumulated charge. The body that is to be discharged is brought close to the body that is to be charged. The closer the two bodies approach, the more charge they contain at a given potential difference. The best results are obtained by using parallel plates very close to each other. The field attracts the charges of the two plates so that they sit only on the two plate surfaces facing each other. The amount of charge on the plates does not change if one of them is grounded, Fig. 19.15. This type of storage device for electric charge is called a *capacitor*.

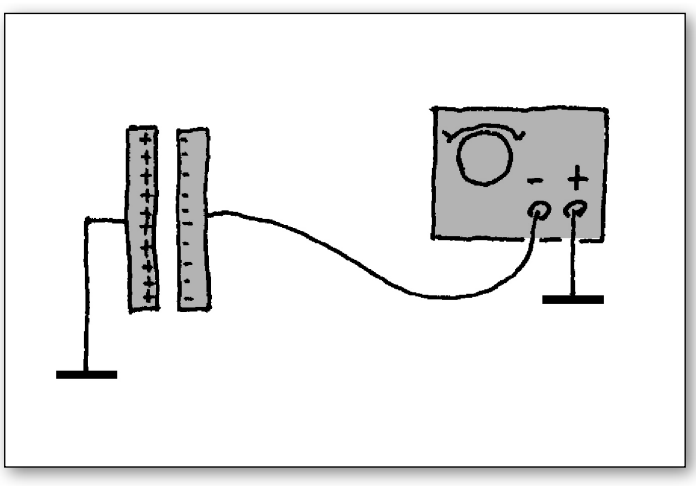

Fig. 19.15
A capacitor. Two metal plates are very close together. The charges on them have opposite signs.

We will do an experiment that shows that the capacity of a capacitor increases when the distance between the plates is decreased. We use a capacitor with adjustable plates for this. We ground one plate, perhaps the negatively charged one. First, we charge the capacitor with the plates far apart and discharge it through a glow lamp. We repeat this several times, each time slightly reducing the power supply's voltage, Fig. 19.16. Finally, the capacitor's charge is so small that the glow lamp shows no noticeable reaction. We now move the plates very close together until they are about 2 mm apart. We repeat the experiment again using the lowest voltage from before. The glow lamp burns strongly now. We find that at the same voltage, there is more charge in plates close together than in ones far apart.

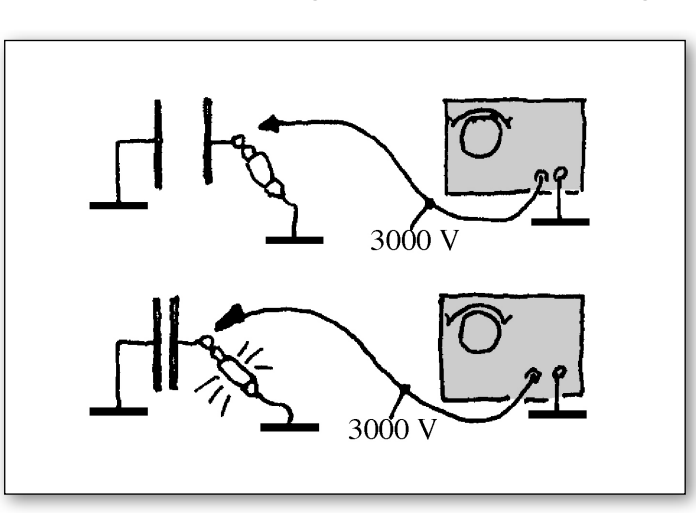

Fig. 19.16
Reducing the distance between two plates results in an increase of charge upon them.

The smaller the distance between the plates of a capacitor, the greater the amount of stored charge.

Technical capacitors have plates that are much larger and much closer together than the ones in our experiment. This can be accomplished by rolling two thin layers of aluminum paper, one of which is covered with a thin layer of insulating material, into a cylinder, Fig. 19.17.

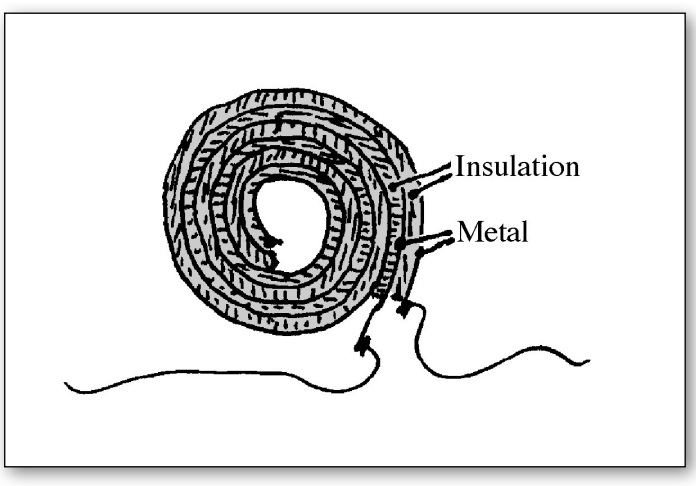

Fig. 19.17
A cross section of a metallized paper capacitor

The charge on a capacitor for technical applications can be so large that it is easy to demonstrate its presence. We construct the circuit of Fig. 19.18. The technical symbol for capacitors is included in the sketch. The capacitor is charged by connecting it to a 6 V power supply for a short time. It is then connected to a small electric motor and the motor runs for a few seconds.

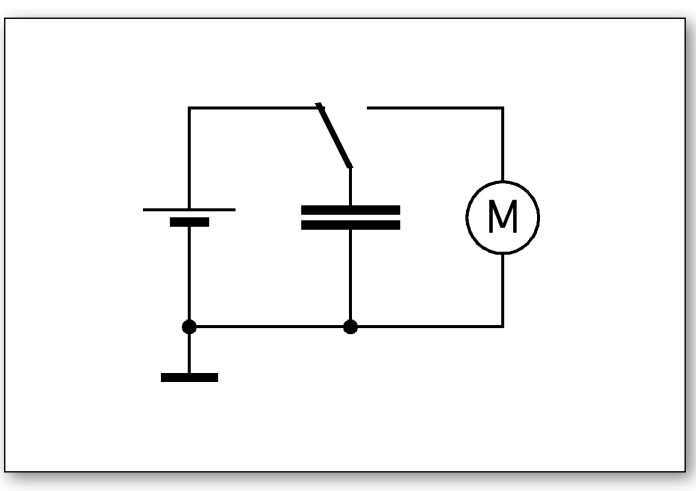

Fig. 19.18
A capacitor as energy storage unit

This experiment shows us a use for capacitors. They can store energy. When the plates are charged with electricity, an electric field forms between them. This contains energy just like a magnetic field does. In the process of charging, energy goes from the power supply into the capacitor. The capacitor gives its energy to the motor while discharging.

A capacitor can have a similar function as an accumulator, i.e., a rechargeable battery. It also stores energy. During charging, it receives energy with the energy carrier electricity. It discharges using the same energy carrier. There are two differences to a rechargeable battery, though. Energy can be brought in and sent out much more quickly by a capacitor than an accumulator can do this. However, the energy capacity of a capacitor is much smaller than that of an accumulator.

Capacitors can be found in all electronic devices.

19.6 Capacitance

We will determine how much electric charge is found on the plates of a capacitor. As we charge the capacitor, the voltage between the plates increases. We continue charging the capacitor until the voltage reaches the highest allowed value. (This value is printed on the capacitor. When it is exceeded, a breakdown can occur between the plates.)

We charge the capacitor with the help of a power supply with a constant current. We use the relation

$Q = I \cdot t$

to obtain the amount of stored electricity.

We know the strength I of the charging current. We measure the time t needed for charging. We can now calculate the amount of electricity Q that is on the plates at the end of the charging process. More precisely, Q represents the charge on the positive and $-Q$ the charge on the negative plate.

Example: The capacitor is charged until it reaches 6 Volts. The electric charge current is 10 mA. Charging takes 6 seconds. The resulting amount of electricity on the plates is

$Q = I \cdot t = 10\ \text{mA} \cdot 6\ \text{s} = 60\ \text{mC}$.

We discharge the capacitor and recharge it once again, but this time to only half the voltage. We find that we need half as long to do this, and the plates have just half the previous amount of electric charge on them. We conclude that the amount of electricity on the plates is proportional to the voltage between them, Fig. 19.19:

$Q \sim U$.

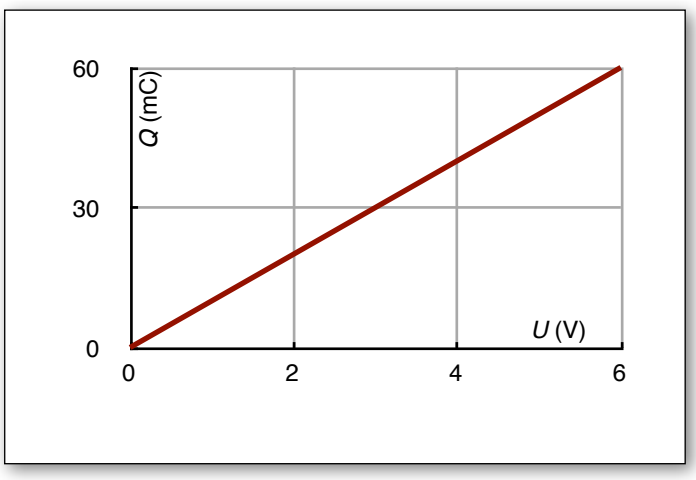

Fig. 19.19
The charge on the plates of a capacitor as a function of the voltage between the plates.

Alternatively, we can obtain this relationship as follows. While the charging is going on, we draw a diagram of the charge on the plates as a function of time. The charge current is constant, so the amount of charge increases uniformly with time. Q is a linear function of time, Fig. 19.20a. During charging we also sketch the voltage between the capacitor plates as a function of time. This results in another linear function, Fig. 19.20b. A comparison of Fig. 19.20a and 19.20b shows that U is a linear function of Q.

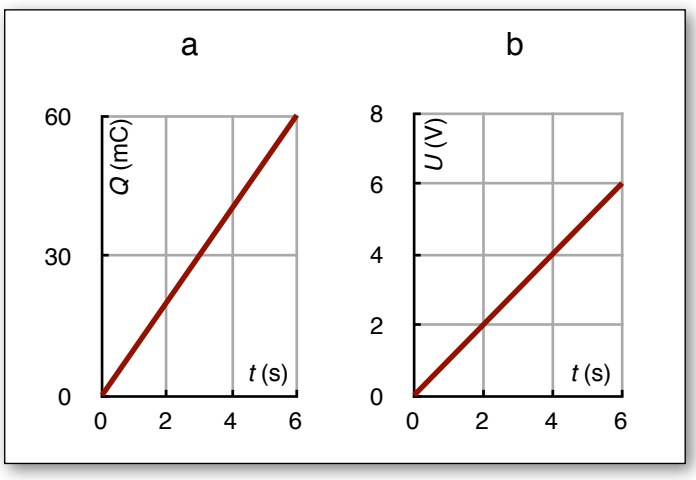

Fig. 19.20
Charge and voltage while charging a capacitor, as a function of time.

Because $Q \sim U$, the quotient Q/U is constant. This quotient is called the capacitance C of the capacitor:

$$C = \frac{Q}{U}.$$

If the capacitance of capacitor A is twice that of capacitor B, at a given voltage, there will be twice the charge on A as compared to B.

Capacitance values are printed on technical capacitors. The equation above shows us that the unit of capacitance is Coulomb/Volt. The abbreviation for this is Farad (F). Therefore

1 C/V = 1 F.

A Farad is a very large unit. Usually the capacitances of technical capacitors are measured in nanofarads or millifarads.

Example: The capacitor in our example had an electric charge of 60 mC when there was 6 V between the capacitor plates. The resulting capacitance of the capacitor is:

$C = Q/U = 60\ \text{mC}/6\ \text{V} = 10\ \text{mF}$.

The voltage U between the plates of a capacitor is proportional to the electric charge Q on the plates. The quotient Q/U is called the capacitance of the capacitor.

Exercises

1. A capacitor is charged by a constant electric current of 2 mA until the voltage between the plates reaches 240 Volts. Charging takes 2 minutes. (a) How much electric charge is on the plates at the end of charging? (b) What is the capacitor's capacitance?

2. There is a voltage of 150 Volts between the plates of a 80 µF capacitor. How much electricity is on the plates?

19.7 Atmospheric electricity

If a grounded cable is put near the charged sphere of a Van de Graaff generator, “a spark flies over”. To be more exact, the following occurs. There is always a small number of ions in the air. These become so strongly accelerated in the field between the end of the grounded cable and the sphere that they ionize other molecules in the air when they collide with them. These new ions ionize more molecules, etc, forming a “tube” of ionized air that is conductive and lights up. The sphere discharges through it.

A thunderstorm shows this phenomenon much more impressively in the form of lightening. The air is ionized here as well, but this time between the Earth and a charged cloud. The voltage is also much higher than in a Van de Graaff generator. It can be many millions of volts.

In the atmosphere, the electricity flows in an odd circuit. In order to understand this circuit and to understand the processes taking place during a thunderstorm, we first need to deal with the atmosphere when the weather is good.

The electric conductivity of air at ground level is very small. It grows with increasing altitude because the number and mobility of ions also increases strongly with altitude. Put in simplified terms, we can say that the atmosphere forms a well conducting layer above an altitude of about 50 km. This part of the atmosphere is called the ionosphere. The air pressure here is less than 1/1000 of that on the Earth’s surface. The ionosphere and the Earth’s surface form what could be considered the two plates of a gigantic capacitor.

The ionosphere is at a constant potential of about 300,000 volts. There is the badly conducting layer of air with a total resistance of about 200 Ω between it and the Earth’s surface. The potential difference between the lower edge of the ionosphere and the surface of the Earth leads to a vertically oriented electric current of 1,500 A, Fig. 19.21.

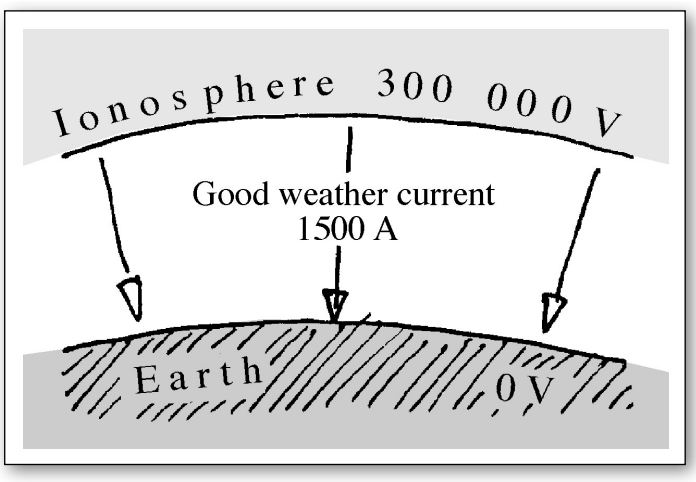

Fig. 19.21
The electric potential of the ionosphere is about 300,000 volts higher than that of the Earth. In good weather, a current of about 1500 A flows from the ionosphere to the Earth.

This fair weather electric current goes almost unnoticed because it is distributed over the entire Earth. The ionosphere would discharge quickly as the result of this current: the voltage would collapse in no more than a half an hour. The reason that it does not so are the thunderclouds.

A thundercloud can be pictured as a huge electricity pump, a kind of battery or a Van de Graaff generator. We will soon see how it functions. At the moment, though, we will only consider how it is built into our atmospheric circuit, Fig. 19.22. The thundercloud pumps electricity upward from below. A strongly negative potential forms on its underside while a strongly positive potential forms on its upper side.

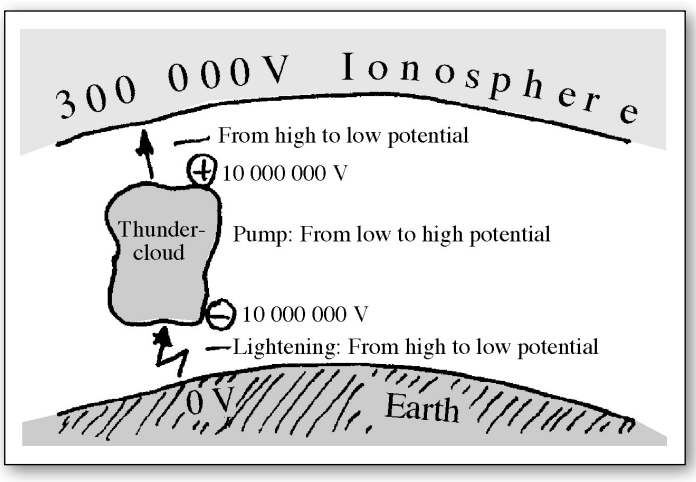

Fig. 19.22
Thunderclouds continuously pump electricity from the Earth to the ionosphere.

Because the air above the cloud conducts fairly well, an electric current flows from the upper ‘terminal’ of the cloud into the ionosphere. It flows from the higher potential (several million volts) to the lower potential (300,000 V).

In addition, a current must flow from the Earth to the lower “terminal” of the cloud. It flows from the higher potential (0 V) to the lower one (minus several million volts). The air under the cloud is a bad conductor, as we know, so this current can only flow over a lightning bolt. Lightning bolts represent the feed line to our electricity pump.

How does this pump work? What happens in a thundercloud? The drops of water and ice particles in a cloud are electrically charged. The smaller particles tend to be positively charged and the larger ones negatively charged. Overall and averaged out over all the particles, the cloud is neutral. However, in a thundercloud, processes occur that separate the large and small particles from each other. When this happens, the positive and negative charges become separated as well. This causes a potential difference to form.

The strong updraft inside the cloud is responsible for the separation of particles according to size and therefore the separation of the positive and negative electric charge. The small particles with their positive charge are carried upward, and the large graupels or hailstones fall downward as precipitation. In the process, the cloud’s potential becomes very strongly positive on its upper side and very strongly negative on its underside.

Charging of the ionosphere by thunderclouds is not a process that just happens now and then. The ionosphere can be considered a large continuous capacitor plate. All the thunderclouds around the Earth contribute to charging it. There are always about 2000 thunderstorms happening around the Earth with a total of about 100 lightening discharges per second. The electricity pump charging the ionosphere runs continuously.

20

Data Systems Technology

20.1 Amplifiers

Energy is necessary for every type of transport. A truck bringing bricks to a building site uses diesel fuel and energy along with it. Pumps are needed for water or crude oil to flow through pipes and these pumps need energy. Where does the energy that is used for these transports go? It always causes the path of the transport to warm up somewhat because entropy is produced. This entropy distributes into the surroundings along with the energy which then becomes unnoticeable.

Energy is also necessary to transport data. In most cases, this energy is sent from the data source together with the data. It is given to the data as a kind of provision for the road. In this way, the sound waves produced by a loudspeaker, the electromagnetic waves from an antenna or the light coming out of a television screen carry not only data but energy as well.

We can now complete the graphic symbol of the data source, Fig. 20.1

Fig. 20.1
The sound waves coming from the loudspeaker carry data and energy.

In this example, the energy and the data have the same carrier.

Some types of data transport are especially wasteful of energy. The energy emitted by the source distributes over a growing area, Fig. 20.2. This is the case for sound coming out of a loudspeaker or from a speaking person, or for electromagnetic waves coming out of a radio or television antenna. This kind of distribution is practical when many receivers should be reached without each one needing to have a cable leading to it.

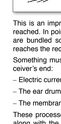

Fig. 20.2
The energy emitted from the source spreads out over an increasingly large area.

This is an impractical method however, if just one receiver is to be reached. In point-to-point radio, for example, electromagnetic waves are bundled so that most of the energy being sent with the data reaches the receiver antenna.

Something must always be "activated" or "triggered" at the data receiver's end:

– Electric currents must be induced in a receiving antenna;
– The ear drum of the person listening must vibrate;
– The membrane of a loudspeaker must be activated.

These processes can only take place when enough energy is sent along with the data. If the energy loss in a telephone cable is too large, or if the distance between a radio transmitter and the receiver is too great, the data transport will fail.

In order to assure that data will arrive with enough energy at the receiver in spite of great distances, a so-called *amplifier* can be used. An amplifier has a data inlet and a data outlet. The data enters the amplifier along with just a little energy and leaves it with a lot of energy. The data current gets new "provisions for the road". Fig. 20.3 shows an electric amplifier symbolically. The data carrier at the inlet and at the outlet is electricity.

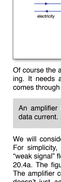

Fig. 20.3
A symbolic representation of an electric amplifier

Of course the amplifier does not create the extra energy out of nothing. It needs a second inlet for this. In many cases, this energy comes through an electric cable into the amplifier.

> An amplifier increases the energy current accompanying the data current.

We will consider an electric amplifier to clarify what amplifiers do. For simplicity, we will assume that the data is in binary code. A "weak signal" flows into the amplifier. It could look like the one in Fig. 20.4a. The figure shows the energy current as a function of time. The amplifier converts it into a "strong signal." It is important that it doesn't just *add* a constant energy current to the weak signal as shown in Fig. 20.4b. The result would still be called a weak signal because the difference between the greater and smaller values

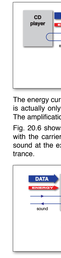

Fig. 20.4
Energy current *P* as a function of time *t*
(a) A weak signal
(b) After adding an energy current of constant strength, the signal is still weak
(c) A strong signal

would be just as hard to discern as in the signal of Fig. 20.4a. The amplifier must *multiply* the energy current by a large factor. The result of multiplying by a factor of 8 is shown in Fig. 20.4c.

An amplifier can be characterized in terms of the ratio of the outgoing to the incoming energy current.

Fig. 20.5 shows the data flow from a CD player to its loudspeakers. The player produces an energy current of about 0.1 μW. The loudspeakers need about 10 W, however, so there is an amplifier between the speakers and the player. The amplification factor in a typical hifi amplifier is about 10^8.

Fig. 20.5
Data transfer from a CD player to the loudspeakers

The energy current of an electric signal coming from a radio antenna is actually only about 1 pW (one millionth of a millionth of a Watt). The amplification factor for radio reception therefore is about 10^{13}.

Fig. 20.6 shows a flow diagram of a megaphone. The data enter it with the carrier sound and leaves it with the same carrier, but the sound at the exit carries much more energy than it does at the entrance.

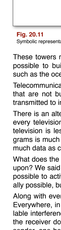

Fig. 20.6
Symbolic representation of a megaphone

A relay is a very simple amplifier for binary digits, Fig. 20.7.

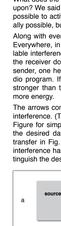

Fig. 20.7
The relay is an amplifier for binary digits.

The corresponding flow diagram is in Fig. 20.8.

Fig. 20.8
Symbolic representation of a relay

This type of amplifier can be easily transformed into an amplifier for optical binary digits, Figs. 20.9 and 20.10.

Fig. 20.9
An amplifier for optical binary signs

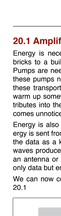

Fig. 20.10
Symbolic representation of the amplifier of Fig. 20.9

Television programs often take a complicated path to the viewer. A television station has a range of only about 50 km. In order to broadcast a television program over a longer distance, substations with amplifiers are needed. These are the so-called *telecommunication towers* that we sometimes see on top of mountains. Data are transmitted from one of these towers to the next with bundled electromagnetic waves. Each tower has an antenna for receiving and one for sending. The signals are amplified between the receiver and the sender, Fig. 20.11. These same stations also serve to transmit telephone conversations over longer distances.

Fig. 20.11
Symbolic representation of the electric system of a telecommunications tower

These towers must always be within sight of each other. If it is impossible to build a tower for transmission across large distances such as the ocean, the station can be put into a satellite.

Telecommunication towers also have sender antennas for waves that are not bundled. This is how television programming can be transmitted to individual television sets.

There is an alternative for this last step of data distribution by which every television set receives data. This is cable distribution. Cable television is less susceptible to disturbance and the choice of programs is much larger. Wireless data currents cannot carry nearly as much data as cables can.

What does the amount of energy needed by a data receiver depend upon? We said that something must be activated in the receiver. Is it possible to activate something using very little energy? This is actually possible, but it poses new difficulties.

Along with every actual data source there are other "data sources". Everywhere, in every line and every data transfer device, uncontrollable interference occurs. Interference is also data – but data which the receiver doesn't want. If a radio is tuned to a weak or far away sender, one hears all kinds of interference along with the actual radio program. If data reception is good, the desired signal is much stronger than this noise. In order to be stronger, it needs to have more energy.

The arrows coming in from below in Figs. 20.12a and 20.12b show interference. (The data and energy carriers have been left out of the Figure for simplicity). Fig. 20.12a shows bad data transfer because the desired data has less energy than the interference. The data transfer in Fig. 20.12b is good. Compared to the desired data, the interference has very little energy. In this case, the receiver can distinguish the desired data from the interference.

Fig. 20.12
Symbolic representation of data transfer with interference.
(a) Bad transmission.
(b) Good transmission

There is a game that shows the results when a receiver cannot distinguish between the desired data and interference. In this game several people stand in a row and the first person whispers a sentence into the ear of the person next to her. She then whispers this sentence, or what she believes she heard, to the next one in line, etc. Finally, the last one in line often hears something that has nothing at all to do with the first sentence. If all the players had spoken aloud (given their data more energy) the original sentence would probably have reached the last receiver unchanged.

Exercises

1. Name four examples of data transfers that use amplifiers.
2. A person can play the role of an amplifier. Give an example.

20.2 Data processing

We have already looked into data transport and storage. Now we will turn to data processing. This takes place in natural as well as technical systems. A natural system where data processing occurs is the human brain or the brain of an animal. A technical system is the computer. Data processing systems have a hierarchical structure. Their functionality can be described at different levels of this hierarchy. First, we will describe the general structure, as if we were looking at it from a distance.

The general structure of data processing systems

A computer system consists of the following elements: the computer itself, several input and output devices, and one or more external storage units, Fig. 20.13.

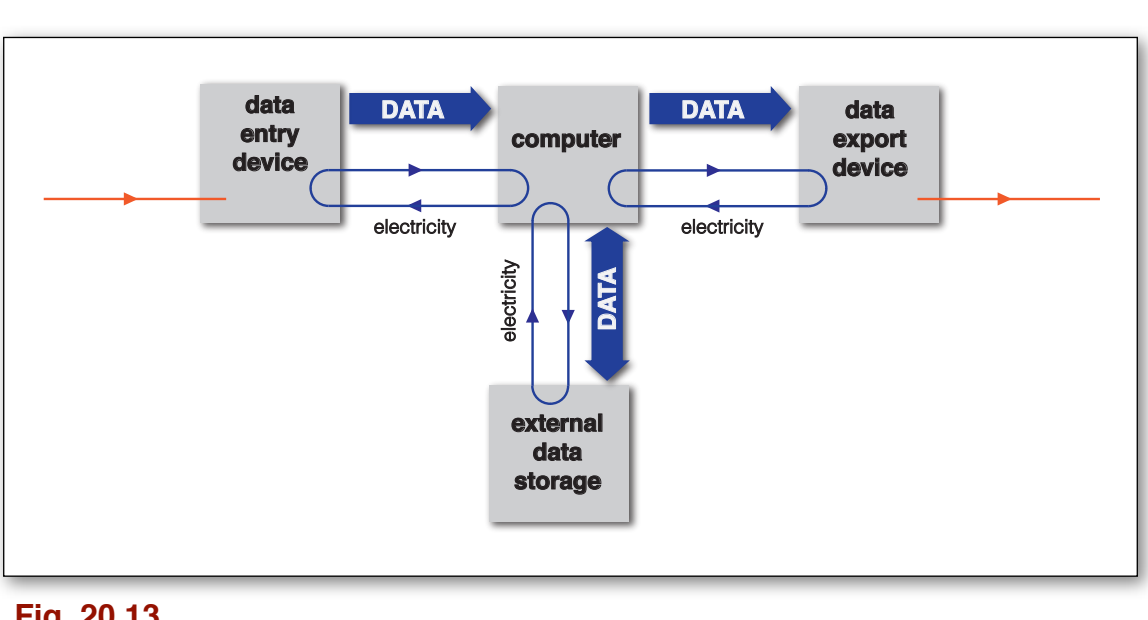

Fig. 20.13
Configuration of a computer system

An input device serves to prepare incoming data for the computer by transferring it from the carrier it comes in on to the carrier electricity. Also, the input device converts data into a code the computer can work with.

The data output device converts the data coming out of the computer into a code that the user can understand. "Users" can also be machines controlled by the computer.

There are several kinds of input and output devices, Table 20.1.

Input devices
Keyboard
Scanner
Barcode reader
Joystick
Mouse
Video camera
Microphone
Various sensors
Output devices
Monitor
Printer
Loudspeaker
Various actuator

Table 20.1
A computer's input and output devices

Most computers have a keyboard as input device. The letter signs are first put into a seven bit code and then further processed.

Other well known input devices are the joystick and the mouse.

Bar code readers can be found at the cash registers in department stores and supermarkets. Each item has a field of bars printed on it. This is a coded series of digits. The bar code reader reads the width of and distances between the bars and determines what number the bar code corresponds to. This number helps the computer identify the item. It then tells the cash register the price. In addition, the warehouse finds out how many have been sold of which item in the store.

In order to process pictures or sound, a computer needs a camera, a scanner or a microphone as input device.

The most familiar output devices are monitors, printers and loudspeakers.

Computers are often used to control industrial processes, for example, in chemical factories and power plants. So-called *sensors* are the input devices for such computers. A sensor is a device that measures some physical quantity, possibly temperature, pressure or the concentration level of some substance. It then passes the data on with the carrier electricity. The computer then calculates, based on these values, whether and how the process in question should be influenced. It then gives the data representing the results of its calculations to the output devices, which are called *actuators*. Examples of actuators are the electrically controlled valves or switches that turn pumps, ventilators, heaters or similar devices, on and off, in order to influence the processes taking place in the factory.

The third component of the computer system is the *external storage unit* (Fig. 20.13). Although there is a data storage unit inside the computer itself, it is often not large enough for some applications. This is why there are external storage units connected to it.

Natural data processing systems have similar structures. The computer in Fig. 20.13 corresponds to the brain, the input devices correspond to sense organs and the voice or maybe the writing hand act as the output device. A person also uses "external storage units" when he believes he cannot rely upon his memory (his "working memory"). Notebooks, address books, libraries, etc. are examples of these external storage units.

The lower levels of data processing systems

We will now continue with the description of data processing systems at a much lower level. The structures we are dealing with here are so small that they can only be seen with a microscope.

At this level of a computer, we find the electronic components. The transistors are the most important of these. Later on, you will get to know how these work. At this point we will only say that a transistor is an electrically controlled switch. It does what a relay does, only much faster. Fig. 20.14 shows the symbol for a transistor. Terminals 1 and 2 are input and output for the electric current. This current can be turned on and off with the help of terminal 3. Depending upon the value of the electric potential of terminal 3, the electricity will or won't be allowed to flow between terminals 1 and 2.

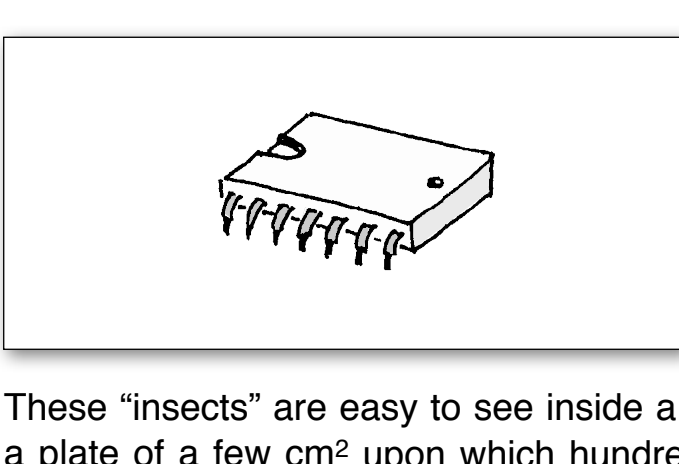

Fig. 20.14
Symbol of a transistor

All the data processing in a computer can be reduced to the opening or closing of such "switches". A switch can be either open or closed, it has *two* states. Therefore, the data processing in a computer must be done in binary code.

We will now move up one level in our description of the computer. We are now on the level of the *chip*. A chip is an element combining many transistors. Every chip is built into a small black plastic case. This case has many terminals protruding in two rows from the case like the feet of an insect, Fig. 20.15.

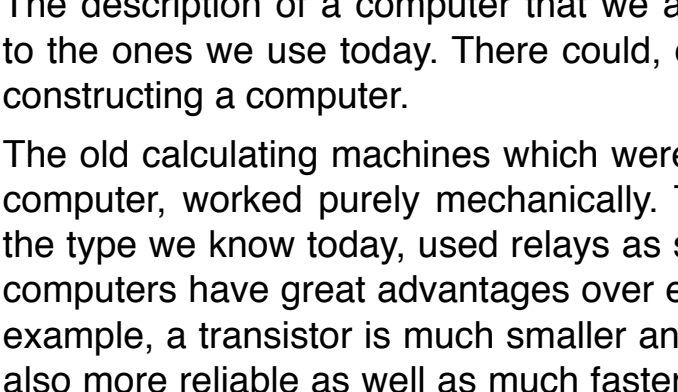

Fig. 20.15
The outside of a chip

These "insects" are easy to see inside a computer. The chip itself is a plate of a few cm^2 upon which hundreds of thousands of transistors are arranged in two dimensions. Some chips have diodes and resistors as well. The most important components, however, are the transistors. These transistors (along with the diodes and resistors) are connected to each other and make up a complicated "control unit" so that the chip can carry out certain operations. Some special functions for chips would be storage, calculation, coding, timing…

At a later point we will look at how it is possible to carry out logical operations by interconnecting transistors.

The description of a computer that we are giving here corresponds to the ones we use today. There could, of course, be other ways of constructing a computer.

The old calculating machines which were already a kind of primitive computer, worked purely mechanically. The very first computers of the type we know today, used relays as switches. Modern electronic computers have great advantages over electromechanical ones. For example, a transistor is much smaller and cheaper than a relay. It is also more reliable as well as much faster. The fact that an electronic computer has many advantages, does not mean that a totally different and much more powerful type of computer couldn't be developed. The idea of optical computers is being researched. It is expected that they will work much faster than the electronic computers of today. The data carrier would not be electricity, but light.

The brain is made up of many very tiny components each of which can be in one of two states. These are the nerve cells, so-called *neurons*. There are about 10^{10} neurons in the brain. This is more than 100 times the amount of transistors in a supercomputer. Moreover, a nerve cell is a considerably more complex component than a transistor. One neuron has about 10,000 inlets but only one outlet, Fig. 20.16. The conduit at the outlet branches off greatly and is connected to the inlets of other nerve cells. This structure is called a *neural network*.

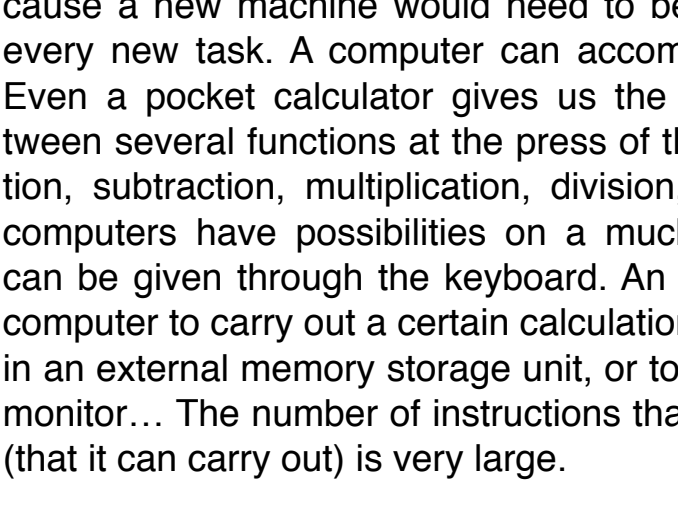

Fig. 20.16
Symbolic representation of a neuron

Each inlet of a nerve cell can be at a high or low electric potential. Whether the outlet of a nerve cell is at a high or low potential depends upon the number of inlets that are at high potential.

Computer programs

It is easy to imagine a data processing machine that has only one type of job to do. Such machines would be impractical, though, because a new machine would need to be constructed or bought for every new task. A computer can accomplish many different tasks. Even a pocket calculator gives us the possibility of choosing between several functions at the press of the appropriate button: addition, subtraction, multiplication, division, root extraction, etc. Real computers have possibilities on a much larger scale. *Instructions* can be given through the keyboard. An instruction might be for the computer to carry out a certain calculation, to read a numerical value in an external memory storage unit, or to write a certain letter on the monitor… The number of instructions that a computer "understands" (that it can carry out) is very large.

Data reduction

A data processor receives data and transmits data. This is also true of data transporters and coders. Does this mean that a computer is nothing more than a coding machine? That the the data at the inlet carry the same information as that at the outlet, just coded differently? If this were so, the amount of data coming in would have to be exactly the same as the amount of data going out. We will see if this is true by using a simple example.

A computer is programmed so that it can calculate the average values of numbers. For simplicity, we will assume that the numbers being entered are integers and that the average value is calculated as an integer as well.

We have our computer calculate the average grade of a test. We assume the class has 30 students, and that a maximum of 15 points can be achieved on the test. 30 numbers are entered into the computer of which each is one of the 16 different integers from 0 to 15. Because $16 = 2^4$, the computer gets 4 bit with each number, in total

$30 \cdot 4$ bit = 120 bit.

The program starts up and in a short while, the computer produces the average grade: one of the numbers from 0 to 15. The amount of data is 4 bit, Fig. 20.17. Thus, the computer has "reduced" the amount of data.

> Computers reduce the amount of data.

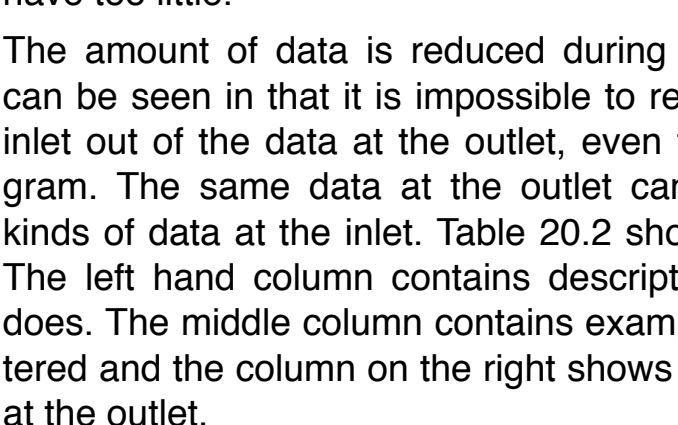

Fig. 20.17
The amount of data going into the computer is greater than the amount of data coming out of it.

This is surprising. Does it mean that someone receiving the data at the outlet knows less than someone receiving data at the inlet? Yes, this is exactly what it means. The person who knows the average grade cannot reconstruct the students' individual grades, but is this necessarily bad? Why do we use computers at all? They are used because it is very difficult to deal with all the data at the inlet. The amount of data at the inlet would be too large to compare this class with a parallel class. It is easy to lose the overview when there is a lot of information. The amount of data is too much to deal with. People use computers when they have too much data, not when they have too little.

The amount of data is reduced during computer processing. This can be seen in that it is impossible to retrieve data that was at the inlet out of the data at the outlet, even though one knows the program. The same data at the outlet can be produced by different kinds of data at the inlet. Table 20.2 shows some examples of this. The left hand column contains descriptions of what the computer does. The middle column contains examples of data that can be entered and the column on the right shows the data as it would appear at the outlet.

What the programmed computer does	Input data	Output data
Addition of positive and negative integers	10; 5	15
	5; 10	15
	14; 1	15
	–123; 138	15
Carrying out the AND operation	0; 0	0
	1; 0	0
	0;1	0
Alphabetizing names	Bob; Willy; Lilly	Bob; Lilly; Willy
	Bob; Lilly; Willy	Bob; Lilly; Willy
	Willy; Bob; Lilly	Bob; Lilly; Willy
Rounding out the numbers to three digits	2,7184	2,718
	2,7182818	2,718
	2,7176	2,718
Calculating the square root to three digits	2	1,414
	1,998	1,414
	2,0007	1,414

Table 20.2
The amount of output data is smaller than the amount of input data.

The two last examples show that just rounding a number leads to data loss.

In some cases, the amount of data is exactly as great at the outlet as it was at the inlet. This means that the input can be retrieved from the output. A simple example is reversal of algebraic signs. Any coder has also this property. If a computer is used only to store data, there will be no data reduction of data either.

We will perform a balance of the amount of data for a more complicated example. Now the computer will be used for image recognition, Fig. 20.18. A digital camera is the input device and a printer is the output device. Ideally, the image recognition process would go like this: The camera is pointed at the object to be recognized, maybe a rose. The program starts up and shortly thereafter, the printer prints the word "rose". Practical applications of this procedure are reading of handwriting or fingerprint recognition.

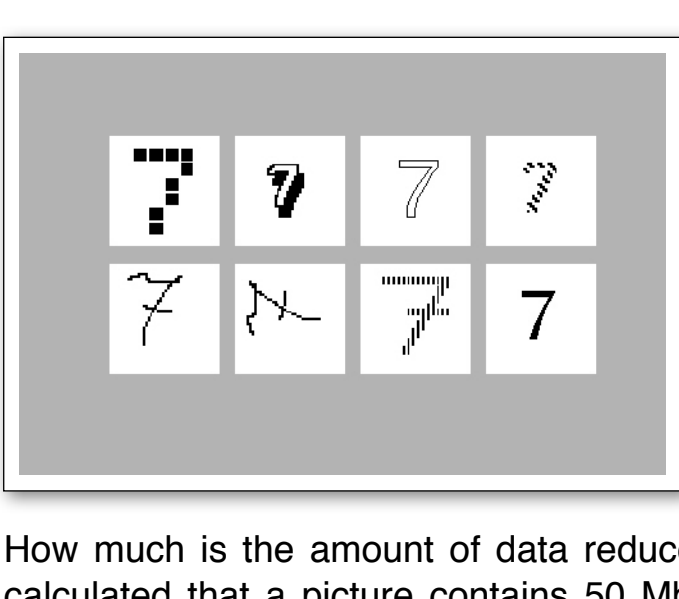

Fig. 20.18
The image recognition system is made up of a video camera, a computer and a printer.

We have a sheet of paper with the number seven written on it. The seven can be written in many different ways, Fig. 20.19. No matter how it looks, the printer prints the same thing: a plain seven of its own font. There can be many different images at the inlet – many different ways of writing seven – but the same image always appears at the output device.

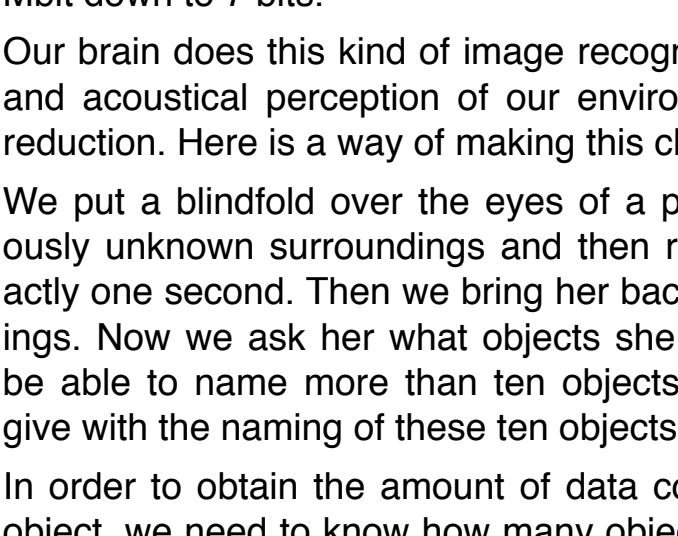

Fig. 20.19
The computer should recognize each of these images as the number seven.

How much is the amount of data reduced in this case? Earlier we calculated that a picture contains 50 Mbit. The computer therefore receives 50 Mbit from the camera.

The printer receives instructions from the computer to print a character. A character has 7 bits, as we know.

In this case, the computer has reduced the amount of data from 50 Mbit down to 7 bits.

Our brain does this kind of image recognition every second. Optical and acoustical perception of our environment is based upon data reduction. Here is a way of making this clear:

We put a blindfold over the eyes of a person, bring her into previously unknown surroundings and then remove the blindfold for exactly one second. Then we bring her back into the original surroundings. Now we ask her what objects she saw. She will probably not be able to name more than ten objects. How many bits does she give with the naming of these ten objects?

In order to obtain the amount of data contained in the name of an object, we need to know how many objects with different names exist in all. Every object has a name and it is in the dictionary. A standard dictionary contains 60,000 words. We estimate very roughly that there are 60,000 objects with different names. This means that with each object name given, we receive 16 bits of data. Naming of 10 objects gives us the amount of data

$10 \cdot 16$ bit = 160 bit.

How many bits did the person take in during that one second glance? We have already estimated that a television broadcasts 1000 Mbit per second. Our example here deals with optical perception just like in the case of television. This means that our test person received 1000 Mbit during the one second she was allowed to see.

Here, the amount of data gets reduced from 1000 Mbit to 160 bit. A huge data reduction occurs in the human brain. This is a basic characteristic of what is called perception.

> Perception is based upon data reduction.

The huge current of data flowing through the pupils into the eyes could not be further processed by the brain without reduction.

Is the computer superior to the human?

This is a badly posed question. Of course there are some ways in which the computer is superior—but in others it is not. This is even true of a pocket calculator. It is also true of a car engine or an elephant: They are both stronger than a person. A better question would be: In what ways is the computer superior to a human being? It is extremely superior in making simple arithmetic operations. A normal PC can carry out several millions of simple arithmetic operations per second. There are, however, lots of "data processing" problems that any person can solve much more quickly and better than the largest supercomputers of today. One problem of this sort is optical perception. Even a very well programmed computer has difficulty distinguishing between a photo of a dog and one of a cat or a horse. It is the network structure of the brain that allows a person or an animal to do this easily. There are efforts being made to develop computers that are constructed like neural networks.

Exercises

1. Give reasons why a computer reduces data when it is programmed to calculate the value x^2 from every whole number x and then to show it on the monitor.

2. Is the amount of data reduced when the value of x^3 is calculated? (x = an integer)

20.3 Generalizing the definition of the amount of data

Fig. 20.20 shows three examples of a data transfer between person A (the source) and person B (the receiver). The three examples differ in their sign sets. In (a), there are 2 possible signs, green light or red light. In (b), there are 8 signs, the numbers 1 through 8. Finally, in (c), there are 32 characters: 26 capital letters and 6 punctuation marks. We now ask how difficult it is for the receiver B to guess the next sign coming in before it gets there.

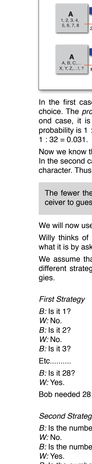

Fig. 20.20
Three data transfers with different character sets

In the first case, the chances are 1 : 2 that B will make the right choice. The *probability* that B will guess correctly is 0.5. In the second case, it is less likely that B will guess correctly because the probability is 1 : 8 = 0.125. In the third case the probability is only 1 : 32 = 0.031.

Now we know that in the first case, 1 bit is transferred with one sign. In the second case, it is 3 bits and in the third example it is 5 bits per character. Thus, the following rule holds:

> The fewer the bits carried by a sign, the easier it is for the receiver to guess what it will be.

We will now use this to analyze the following game.

Willy thinks of an integer between 1 and 64. Bob should find out what it is by asking Willy as few yes-no questions as possible.

We assume that Willy is thinking of number 28. Bob can now use different strategies to find it out. We will compare two such strategies.

First Strategy

B: Is it 1?
W: No.
B: Is it 2?
W: No.
B: Is it 3?

Etc.....

B: Is it 28?
W: Yes.

Bob needed 28 questions to find out the number.

Second Strategy

B: Is the number greater than 32?
W: No.
B: Is the number greater than 16?
W: Yes.
B: Is the number greater than 24?
W: Yes.
B: Is it greater than 28?
W: No.
B: Is it greater than 26?
W: Yes.
B: Is the number greater than 27?
W: Yes.

Bob knows the number after only 6 questions.

This example poses a problem. We had previously defined a bit as the amount of data transferred to answer to a yes-no question. How many bits did Bob get? 28 or 6? The total amount of data transferred cannot depend upon whether the partner is acting intelligently or not. Whatever the strategy, at the end Bob knows the number, meaning the data were received by him.

We can determine how many bits Bob really received by using another more dependable way. Willy could have just told Bob the number to begin with. If he had, it would be easy to give the number of bits. The number represents a choice out of a possible 64 signs. Because 64 = 2^6, the number has 6 bits. Bob needs to receive 6 bits of data. This means that when he used the first inefficient strategy, he received less than 1 bit with each yes-no question.

The fact that Bob received fewer bits per answer with the bad strategy than with the good one, agrees with the rule we found at the beginning of this section: "The fewer bits a sign has, the easier it is for the receiver to guess what it will be." In fact Bob did have a good chance of guessing the right answer by using the bad strategy. He knew that when he asked the question "Is it 1?" he would most probably get a "No" for his answer. The probability of his being wrong was 1 : 64. The probability of his being right was 63 : 64. A good strategy would offer him much less certainty about the answer. No matter whether he assumes that the next answer will be a "yes" or a "no", the probability that he is correct is only 1 : 2.

We conclude from our comparison of the two strategies that one receives 1 bit with the answer to a yes-no question only when both answers are equally probable.

> 1 bit will be transferred with a binary sign only when both signs are equally probable. In all other cases, less than one bit will be transferred.

This is also true when there are more possibilities, i.e., when binary signs are not used. If there are 4 possible signs, one of these will contain 2 bits only when all 4 signs have the same probability. If there is a character set of 8 signs, then each sign contains 3 bits only when all 8 signs are equally probable.

We previously calculated that just about 7 bits are transferred with a letter sign. We silently assumed that all the letters were equally probable. However, in a standard text, not all letters are equally probable. An "e", a "t" or a space appear more often than a "q", an "x" or an exclamation mark. This means that we have overestimated the number of bits.

Whenever a message is transmitted with a character set whose signs are not equally probable, the corresponding code is said to be *redundant*.

For example, the following strongly redundant code could be agreed upon. A message is transmitted with characters, but each letter is transmitted twice. The word "tree" then looks like this: "TTreeee". After the first T has been received, the other letters have unequal probability. The next sign will certainly not be an a, or A, or b, or B, etc. The receiver knows very well to expect another T. He also knows that after the first r has arrived, the next letter will be an r as well.

Redundancy increases the cost of the transmission. The transmission takes longer. In spite of this, redundancy is often wished for because the message in a redundant code is less vulnerable to interference during transmission. Even if the character series TTreeee loses a few characters along the way, the word T*r*ee can still be recognized by the receiver.

Football and the lottery

Willy and Lily have again planned to transmit data with red and greed light signals. Lily should let Willy know

- at 10 o'clock whether the local football team won (win: "green", no win or a tie: "red");
- at 10:05 if Willy has 6 right in the lottery (green means he won and red means he did not).

Which data transfer has more data? To answer this question we use the rule: "The easier it is for the data receiver to guess a character, the fewer bits the character has."

In the first transfer, it is hard to predict whether "red" or "green" will come because the football team has a good chance of winning, but it's opponent is also strong. In the second case, Willy is pretty certain that his lottery game will go like it always has and that he will most probably not win anything. The amount of data transferred in the first case is greater than in the second.

Using our improved definition of amount of data, we come to the same conclusion: We assume that the probability of the local team winning equals exactly 0.5. The probability of the first signal being "green" is exactly that of it being "red". Therefore, according to our improved definition, the first character will transfer 1 bit. In the second transmission, on the other hand, the probabilities of "red" or "green" are very different. "Red" is much more probable than "green". It will therefore transmit less than one bit.

The best way to weigh

Among 27 balls that look alike, there is one that is heavier than the other 26 identically heavy ones. We should find out with the fewest possible number of weightings, which one it is. Only balls may be put on the balance scale, no weights or anything else, Fig. 20.21.

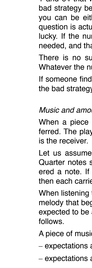

Fig. 20.21
How many weightings need to be performed so we can identify the heavy ball?

Each weighting of the scale answers a question asked of it. The scale can give three different answers: 1, the scale on the right goes down or 2, the scale on the left goes down or 3, equilibrium.

Now if the least possible weightings is the goal, the questions must be asked so that the maximum amount of bits possible are received from each weighting. This means that the three answers must be as equally probable as possible at each weighting. It is surely not very clever to start by putting one ball on each of the scales. The probability of the scale staying in balance is much greater than the probability of the left or right sides moving downward.

How many weightings are necessary?

What is the best strategy?

Good luck and bad luck

We will compare the two strategies for guessing a number between 1 and 64 that we investigated at the beginning of this section. The bad strategy began with the question "Is it 1?" It is easy to see that you can be either lucky or unlucky in this case. If the number in question is actually 1, it was found with only one question. That was lucky. If the number is actually 64, though, 62 more questions are needed, and that is bad luck.

There is no such thing as good or bad luck in a good strategy. Whatever the number is, it will be found out in 6 questions.

If someone finds out the number after one question in spite of using the bad strategy, it can be said that he had more luck than brains.

Music and amount of data

When a piece of music is played and heard, data is being transferred. The player with the instrument is the source and the listener is the receiver.

Let us assume that the instrument is a xylophone with 15 notes. Quarter notes should be made where a quarter rest is also considered a note. If all the notes have equal probability of being made, then each carries 4 bits. If this is not the case, then they carry fewer.

When listening to music, one expects what the next note will be. In a melody that begins like the one in Fig. 20.22, the next note could be expected to be a c. It is not probable that, for instance, an f^1 or an h^1 follows.

A piece of music will be considered unattractive if

- expectations are too often disappointed;
- expectations are too often fulfilled.

Fig. 20.22
What is the next note?

If a melody never continues like expected, it seems chaotic, or "baffling". This is the case when each note is just as probable as all the others or, in other words, when the notes have the maximum number of bits.

However, if the series of notes frequently goes just as expected, the music will seem boring. In this case, each new note transfers so few bits that it is easy to predict.

We have found a rule for composing: The amount of data may not be too large or too small.

Historically, music has developed so that the amount of data has continuously increased. This explains why the modern music of any era has always been found harder to understand than older music that came before.

Exercises

1. Willy is playing with a standard dice (having numbers from 1 to 6). Lilly is trying to guess the number that Willy has thrown by asking as few yes-no questions as possible. What should Lilly pose as the first question so that she receives 1 bit of data with the answer? Give two possibilities. Give the reasons why Lilly would receive less than 1 bit of data with the answer to the question "Is it six?"

2. A card game consists of 32 different cards. What is the minimum number of yes-no questions necessary to find out what a randomly chosen card is?

3. Person A thinks of a term. Person B must try to find out what this term is by asking the smallest number of yes-no questions possible. What strategy should B follow? About how many questions are necessary in this strategy?

21

Light

21.1 Light sources

Objects that emit light are called *light sources*. Some of these are:

– the Sun
– the filament of a light bulb
– a fluorescent lamp
– the flame of a candle
– a light-emitting diode (to be found as an indicator lamp in many electric devices)
– a television screen
– a laser.

An object can be made to glow by heating it up. Any object and any material will begin to glow when its temperature reaches about 800 °C. Some of the light sources listed above (the Sun, fixed stars, or a light bulb) function on this principle. A candle flame glows simply because there are tiny red-hot carbon particles in the flame.

Light can also be created without heating anything up. Fluorescent tubes, light-emitting diodes, television screens and lasers are cold light sources.

Not every object that sends out light is actually a light source. Many bodies only send out light because they receive light. They simply throw the incoming light back, or at least some of it. Most bodies around us belong to this category. There are also celestial bodies, such as the moon and the planets, that do not shine by themselves but reflect light they receive from the Sun.

21.2 Some characteristics of light

Light is a substance, albeit a rather peculiar substance. We will discuss some of its characteristics.

The speed of light

We isolate one ray out of the light being emitted by a light bulb, although it might be easier to just use a laser because lasers create a thin beam of light from the outset.

A spot of light can be seen on the wall where the beam hits it. We interrupt this light beam for a moment by crossing it with our hand. The spot on the wall disappears at the exact moment the hand blocks the beam and reappears as soon as the hand is removed from the beam. The light seems to need no time at all to get from where the hand is to the wall. Actually, it does need a certain amount of time, but it is very, very short. Light moves very fast. It moves at the speed

$v = 300\,000$ km/s.

How is it possible to state this so explicitly? Doesn't the speed of light depend upon how fast it is thrown off its source? Are there light sources that emit slower light? It is possible to create a faster or slower water flow, so can this also be done with light? No, it is not possible to make light faster or slower than 300,000 km/s – at least not when the light is in the air or a vacuum.

Although it is not easy to measure such high velocities, it is possible. There are numerous methods for doing this. Maybe you have some devices at your school to take such measurements.

Moreover, there is one method for braking light in order to make it move more slowly. One lets it run through glass or other transparent solids or liquids. The speed of light in a glass is about 200,000 km/s. In water light moves at about 225,000 km/s. This isn't really braking of the light speed because as soon as it leaves the glass or water, it returns to its usual speed of 300,000 km/s, Fig. 21.1. Every material has a certain light speed belonging to it.

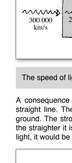

Fig. 21.1
When the light exits the glass, it resumes its original speed.

The speed of light in air or in a vacuum is 300,000 km/s.

A consequence of the high speed of light is that it moves in a straight line. The jet of water in Fig. 21.2 bends down toward the ground. The stronger the jet of water, meaning the faster it moves, the straighter it is. If the water jet could move at the same speed as light, it would be as straight as a beam of light.

Fig. 21.2
The jet of water falls to the ground, but the light beam doesn't (almost).

Light moves in a straight line.

Light is invisible

Is it possible to see light? The question is a contradiction in itself. It is a question of the type: "Can you pay money?" Light is actually an aid we use to see with. When we see an object, light from that object comes into our eyes. We can then say that we "see the object", but not that we "see the light coming from the object". In spite of this, the statement that one cannot see light is hard to grasp. In order to make this a little clearer, we use a laser whose light moves from left to right in our classroom. We see the laser and we see the spot on the wall but we see nothing of the path it takes across the room—unless the air in the room has enough dust floating around in it. Then we can see the path of the beam. Even in this case, however, we are not seeing the light but only dust particles being illuminated.

It is interesting to note that the dark night sky is actually full of light, except for the small area in the Earth's shadow. This light is also invisible to us, Fig. 21.3.

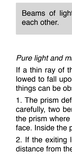

Fig. 21.3
The dark sky at night is full of light.

Even though light is invisible, one often hears about so-called "visible light". This term is used to differentiate it from the kinds of light our eyes have no sensitivity to. Ultraviolet and infrared light are two examples. They are called "invisible" light.

The weight of light

The weight of light is very low, light is very light. (The same word means both things here.)

One could almost believe that light is weightless, and for a long time people believed this to be true. However, it has become possible to determine the mass of light. The light emitted in one hour by a 60 W light bulb weighs about 10^{-13} kg. Our main light source (the Sun) produces light with a very great mass. The light emitted by the Sun per second weighs about four million tons. The Sun loses that much weight per second.

Light penetrating light

We point two light beams onto a wall, Fig. 21.4a. We then turn the two light sources so that the beams of light cross through each other, Fig. 21.4b. What happens if one light beam is now turned off? Is the other one affected in any way? Absolutely nothing happens. Apparently, beams of light can cross through each other with no effect. One simply flows through the other.

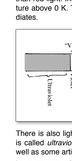

Fig. 21.4
The two light beams do not affect each other.

Beams of light can penetrate each other without influencing each other.

Pure light and mixtures of light – Types of light

If a thin ray of the Sun's white light or the light of a light bulb is allowed to fall upon a prism made of glass, Fig. 21.5, two remarkable things can be observed:

1. The prism deflects the light, bending it. Actually, if one looks very carefully, two bends can be seen. The first one is on the surface of the prism where the light enters, and the other one is at the exit surface. Inside the prism itself, the light moves in a straight path.

2. If the exiting light is allowed to fall upon a white screen at some distance from the prism, the colors of the rainbow can be seen.

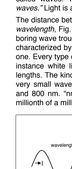

Fig. 21.5
The mixture of light passing through the prism is broken down into its parts.

These observations are interpreted as follows: White light is made up of various components or various types of light. Each of these types of light causes a certain color perception in our eyes. When all of these types of light reach our eyes simultaneously (more exactly: when they all fall upon the same spot of our retina) we see "white".

What is happening with the white light in the prism? A prism deflects light rays but it doesn't deflect the various types of light equally strongly. This is why the white light mixture is decomposed into its parts.

Types of light can be characterized by the color impression they cause in our eyes. Fig. 21.6 shows the types of light arranged according to how strongly they are deflected by a prism.

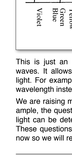

Fig. 21.6
The sequence of types of light on the screen after being broken down by the prism.

Later we will see that color impression is an unreliable indication of types of light because our eyes are made so that they can receive the same color impression in various ways.

What is more, there is light that our eyes do not react to at all. An example would be *infrared* light, Fig. 21.7, which is deflected less than red light. Infrared light is emitted by any object with a temperature above 0 K. The higher the temperature, the more the object radiates.

Fig. 21.7
"Invisible" types of light fall upon the screen. They lie beyond the blue light and the red light.

There is also light that is more strongly deflected than violet light. It is called *ultraviolet* light. It makes up a small portion of sunlight as well as some artificial light sources.

Later on you will learn that, along with the types of light we see, there are many other kinds of radiation you have probably heard of:

– Gamma rays that are emitted by some radioactive substances;
– X- Rays;
– Microwaves;
– Rays used by radar;
– Radio and television waves.

All of these types of radiation are of the same nature. You see that some of them are called waves. Indeed, they all—along with light—have something in common with waves of water, and are therefore called waves. The complete name for them is "*electromagnetic waves.*" Light is an electromagnetic wave.

The distance between two neighboring crests of a wave is called the *wavelength*, Fig. 21.8. (This is also the distance between two neighboring wave troughs, or any other two corresponding points.) Light is characterized by wavelengths, and every sort of light has a different one. Every type of light has its own wavelength. Mixtures of light, for instance white light, are combinations of waves of various wavelengths. The kinds of light that we can see with our eyes have very, very small wavelengths. They fall into the range between 400 nm and 800 nm. "nm" is the abbreviation for nanometer which is one millionth of a millimeter.

Fig. 21.8
The wavelength is the distance between two neighboring wave crests.

Fig. 21.9 shows the relation between color and wavelength.

Fig. 21.9
The relation between color and wavelength

This is just an introductory explanation about the nature of light waves. It allows us, for the moment, to identify different kinds of light. For example, we can say that a laser emits light of a certain wavelength instead of just saying that the light is red.

We are raising more questions here than we are answering. For example, the question arises of how the wavelength of a given type of light can be determined, or what the wave is made up of, exactly. These questions lead us too far away from our actual subject right now so we will return to them later.

21.3 When light meets matter

Light and air influence each other very little. Light goes through air almost as undisturbed as it does through the vacuum between the Sun and Earth. The same holds for other gases.

However, if light hits solid or liquid matter, it can go through great changes. Basically, two things can happen to it:

– it can change direction
– the composition of its components (types of light) can change.

We will begin our investigation by looking into what happens to the direction of light when it hits an object. We need light from a uniform direction for this. It does not need to be any particular kind of light, so we will use a beam of white light.

Reflection and scattering

Many bodies throw back almost all the light that shines upon them. This process can take different forms.

We will have our white light beam fall upon a piece of white paper. The paper sends the light in all directions, Fig. 21.10. The result can be seen in that the entire room becomes bright and all the walls are clear to see. They get their light from where the paper is hit by the beam of light. There is also another way of seeing how the light is reflected in all directions. The point where the light hits the paper is a bright point no matter from which direction it is observed.

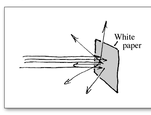

Fig. 21.10
A white matte surface reflects light in all directions, scattering it.

This process is called *scattering*. One would say that the light falling upon the paper is scattered.

We will now let our ray of light fall upon a mirror, Fig. 21.11.

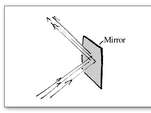

Fig. 21.11
A mirror throws the light back in one direction only. It reflects the light.

Again, the light is completely sent away from the surface of impact – but this time in only one direction producing a single spot somewhere on the wall. Again, there is another way of seeing this: If you look into the mirror, the point where the light hits it is almost invisible except when viewed from a certain direction. This direction is the one into which the light is going away from the mirror. Be careful, though! You must never look in this direction if you are using a laser for the experiment. Laser beams are so strong they will damage your eyes.

What we have just described is called *reflection*.

Transparency and scattering

There are bodies that simply allow light to pass through them. A pane of glass is an example, if the light beam falls perpendicularly upon it.

We insert a pane of glass perpendicularly to the light beam and find that the spot of light on the wall doesn't change in any way, Fig. 21.12. The glass is *transparent*.

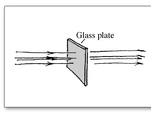

Fig. 21.12
A glass plate allows perpendicular light to flow straight through.

There is a second way to let light pass through a body. A plate of frosted glass lets most of the light through, but scatters it in the process. A part of the light is scattered backwards as well, Fig. 21.13.

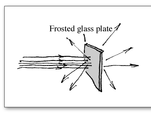

Fig. 21.13
A plate of frosted glass allows light through but scatters it in the process.

Absorption

We now let the light fall upon a piece of black paper. A black woolen pullover would be even better. Almost no light is reflected or scattered from it and none passes through. The light is *absorbed* by the paper or pullover.

More complicated cases

Until now, we have observed special cases: Light has been completely reflected, allowed to completely pass through, or is completely absorbed. Generally, several of these processes happen at once.

A part of light is almost always reflected. Another part is back-scattered, some is allowed to pass through without being scattered. A portion of it is allowed to pass through with scattering and finally, a part is absorbed.

Consider, for example, a sheet of gray glossy paper, possibly the paper used for magazines, Fig. 21.14. The paper reflects a large portion of the light that falls upon it. This reflected light is responsible for the paper's shine. Another part of the light is allowed to pass through and is almost completely scattered. The rest is absorbed by the paper. If no light was absorbed, the paper would be white instead of gray.

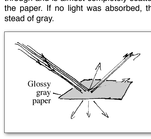

Fig. 21.14
Glossy gray paper reflects, scatters, and allows light to pass through.

This all becomes more complicated yet. What happens to light falling upon an object also depends upon the angle of incidence relative to the object.

Although a glass pane allows most of the light that falls perpendicularly upon it to pass through, it will reflect most light that falls upon it from an acute angle, Fig. 21.15. Try it out.

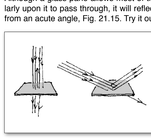

Fig. 21.15
If light falls perpendicularly upon the glass plate, most of it passes through. If it falls upon it at a grazing angle, most of it is reflected.

Everything we have discussed thus far ultimately depends upon the sort of light being used. This is particularly important.

White light (a mixture of all sorts of light) falls upon a tee shirt. The tee shirt absorbs all the types of light except for the blue light, Fig. 21.16. Only blue light is back-scattered by the tee shirt so we say that the tee shirt *is* blue. What would you see if pure red light hit it? Red light would be absorbed and because no other light is available, the tee shirt would appear black.

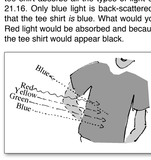

Fig. 21.16
The tee shirt absorbs all the light except for blue. Blue light is back-scattered.

Different sorts of light are back-scattered with varying intensity by different objects, just as some objects allow different kinds of light to pass through to varying degrees. Colored glass is an example of this, Fig. 21.17. The green glass of a soft drink bottle lets green light pass through while absorbing the rest.

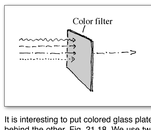

Fig. 21.17
The color filter absorbs all the light except for green. The green light is allowed to pass through.

It is interesting to put colored glass plates (so called *color filters*) one behind the other, Fig. 21.18. We use two color filters that each allow only one different type of light through. When they are put one behind the other, we see that no light is allowed through at all.

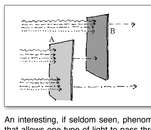

Fig. 21.18
Plate A only allows green light through, plate B only allows yellow light through. When one is put behind the other, they don't allow any light to pass through.

An interesting, if seldom seen, phenomenon results from an object that allows one type of light to pass through and reflects the rest. A pane that does this changes color depending upon the direction it is observed from. It is different when viewed against the light, or from the side the light comes from, Fig. 21.19.

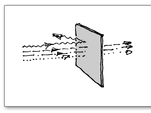

Fig. 21.19
Some color filters reflect the light that they do not allow through.

This has been a general overview of the diversity of phenomena that can be observed when light falls upon an object.

Can you now describe what happens with the sunlight that falls upon a shiny red apple? Can you describe what happens when it falls upon the slightly translucent leaf of a tree?

Exercises

1. Sunlight falls upon a shiny red apple. What happens with the light?

2. Sunlight falls upon the slightly translucent leaves of a tree. What is happening with the light?

3. A dark red pullover looks black when it is lit by blue light of a fluorescent tube. Why is this?

4. There is a kind of packing foil that can be seen through when it is held up to the light. What happens with the light that falls upon this foil?

5. The light of a slide projector falls upon the slide. What happens with the light?

6. A colorful postcard is lit by normal light. What happens with the light falling upon the postcard?

7. The people of a town have built a city hall without windows. In order to light it, they wish to carry in the light in sacks. Why doesn't this work?

21.4 Diffuse and coherent light

We imagine a small spherical space R in the middle of a room, Fig. 21.20. What kinds of light rays pass through this spherical space? Light is coming from very different directions. From the right side, there is a lot of light coming from different parts of the window. Fairly much comes from the walls as well. There is very little light coming from the dark floor below, and somewhere in the room there is a completely black object from which almost no light at all reaches our little space R.

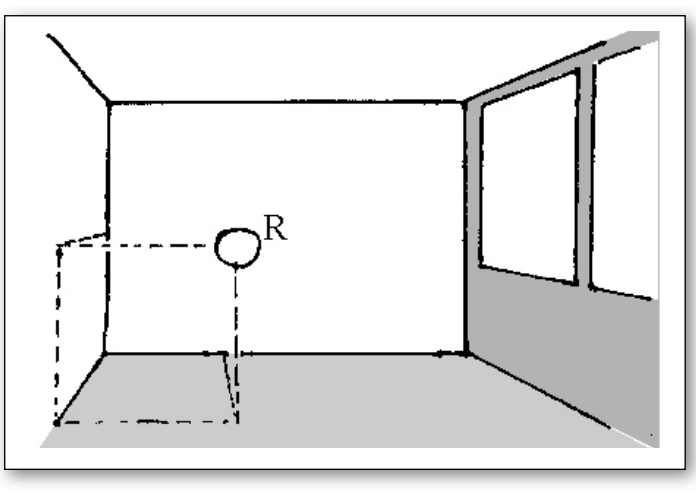

Fig. 21.20
What kind of light passes through the small region of space R?

In order to completely describe the light in R, it is not enough to say how much light travels in each direction. We must also describe how the types of light are distributed. All types of light come from the side of the room with the windows, but only blue light comes into our space from the blue walls.

Let us consider a special situation.

It is a cloudy day. We take our spherical space outside. In Fig. 21.21, R has been magnified and we see a cross section of it. There are a lot of typical light rays drawn in. Different rays of different types of light are drawn with different kinds of lines. We see that light comes in from the right, left, above – in short, from every direction of the "upper half-space". When there is light from different directions *at a point*, one speaks of *diffuse* light.

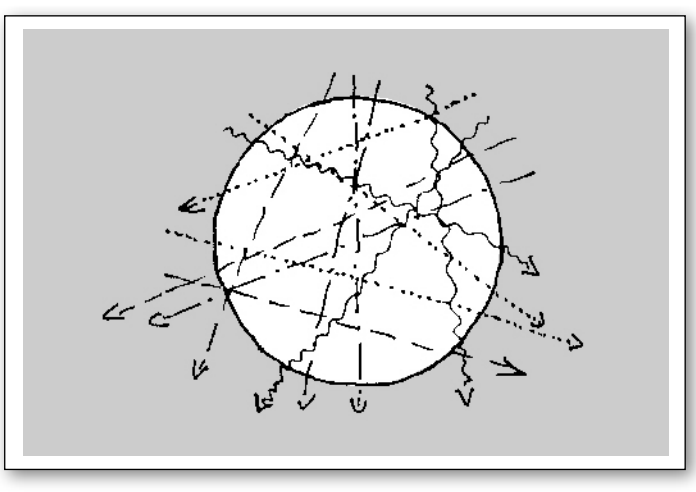

Fig. 21.21
Diffuse light with different colors

Now we imagine a gigantic color filter above the cloud layer. Only one type of light reaches the earth because of this filter. The light that we now would have in our spherical space is represented in Fig. 21.22. It is still diffuse, but in contrast to before, it is now one color or *monochromatic*.

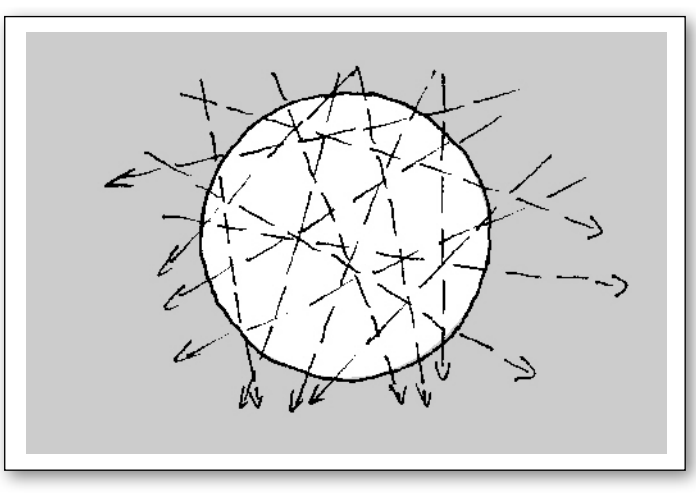

Fig. 21.22
Diffuse monochromatic light

Here is another simple situation. It is night and there is a single light bulb far away from us (and from our space R), Fig. 21.23.

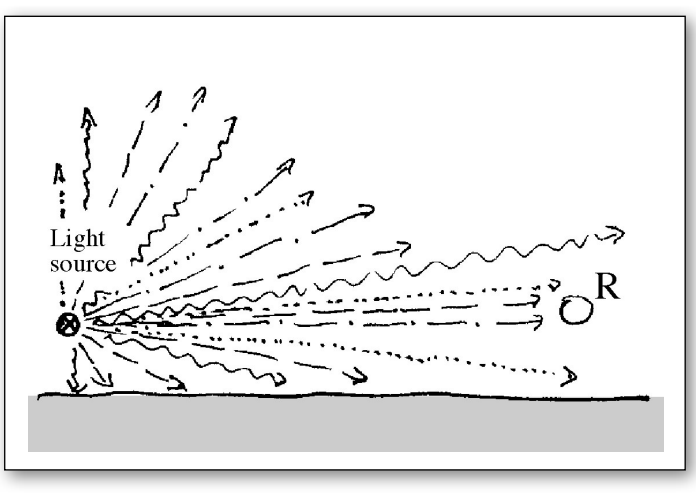

Fig. 21.23
The light in the region of space R has a uniform direction.

Fig. 21.24 shows the distribution of light in R. All the light rays have the same direction.

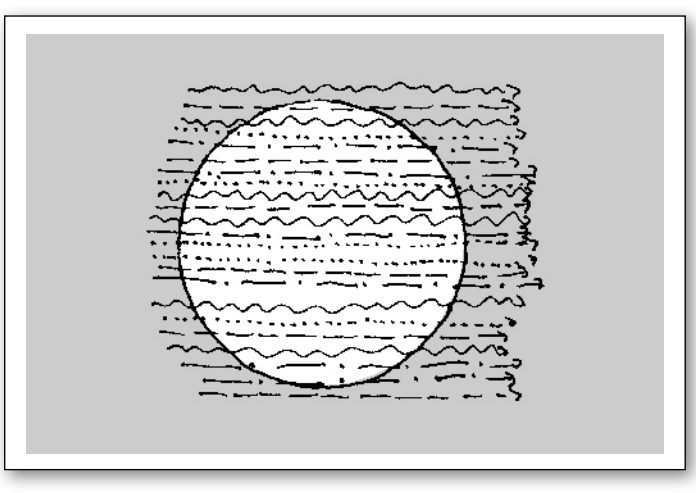

Fig. 21.24
Light with uniform direction

Now we put a color filter in front of the light source. The light shining through R is shown in Fig. 21.25. It is monochromatic and has only one direction. It is the purest light imaginable. It is neither a combination of different types of light as in Fig. 21.24, nor is it a mixture of light from different directions like in Fig. 21.22. It is definitely not a simultaneous combination of different types and different directions as seen in Fig. 21.21.

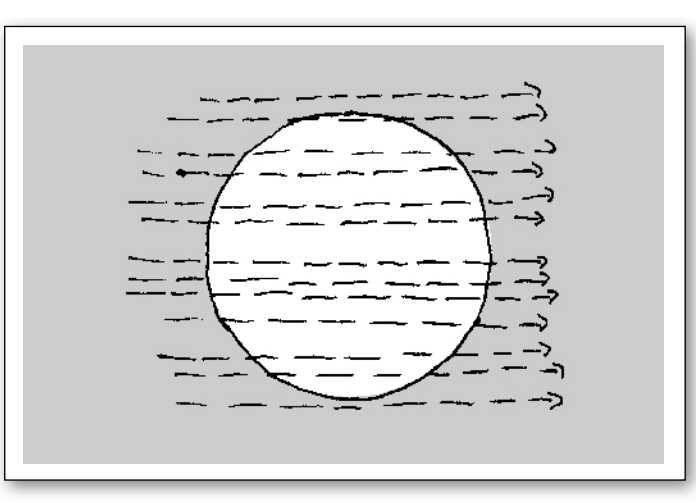

Fig. 21.25
Coherent light: Uniform direction and uniform color

Light that is uniform in type and direction is called *coherent light*. This kind of light is good for experimenting with.

Coherent light: only one type of light with only one direction.

The observations we have just made should make it clear how one would create coherent light. One way would be to put a very small light source with a color filter in front of it very far away. Another way would be to use a closer light source where the "wrong" directions have simply been masked (Fig. 21.26).

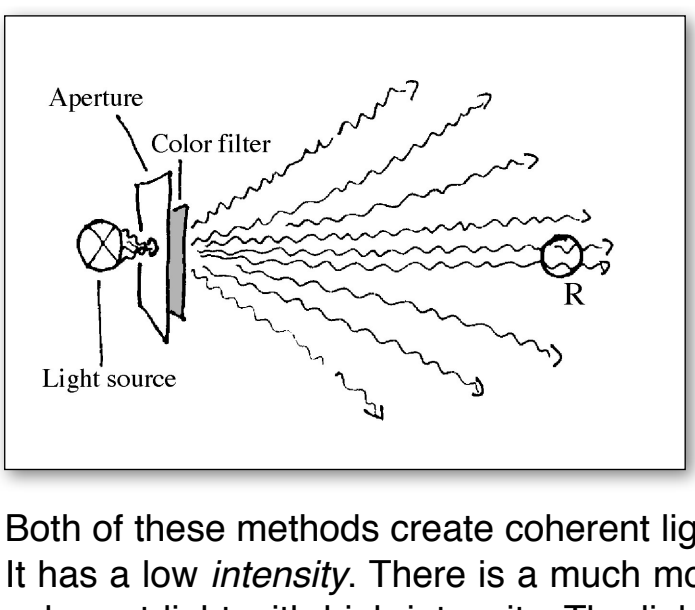

Fig. 21.26
Coherent light can be produced like this.

Both of these methods create coherent light, but it is very weak light. It has a low *intensity*. There is a much more elegant way of creating coherent light with high intensity. The light source for this method is a laser which always creates coherent light. Now we know that the special characteristic of laser light is that:

Laser light is coherent.

Exercises

1. It is such a foggy day that you "can't see the hand in front of your face". How does the light distribution look in a small space in the middle of the fog?

2. It is night and two cars, at great distance from each other, are moving toward an intersection. They approach each other at right angles. What is the light distribution near the intersection (viewed from above)?

3. You are standing on a dark street and see the rear lights of a far away car. What is the light distribution right in front of you?

21.5 Reflection law

We aim a light beam at a mirror. The light is reflected, i.e., deflected in a different direction. What does this direction depend upon? How can it be changed?

It is easy to see that the ray leaves the mirror at a more and more grazing angle the closer the incident light is turned to the mirror. If you measure the *angle of incidence* and the *reflection angle*, you will find that both are the same, Fig. 21.27.

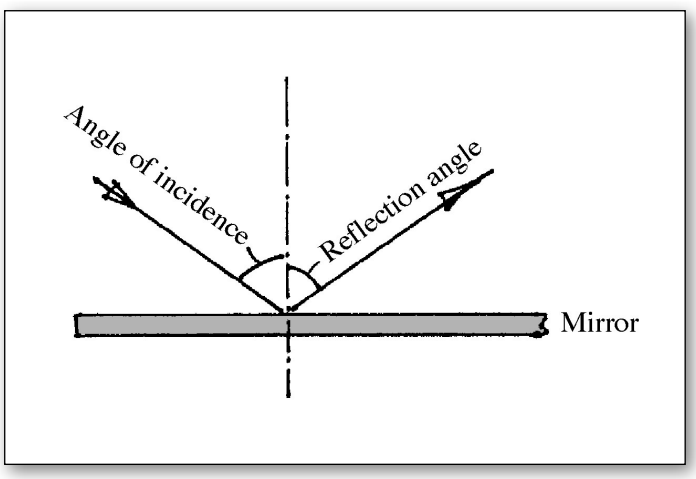

Fig. 21.27
Angle of incidence and reflection angle are the same.

At the same time, it can be seen that both the incident and the reflected rays lie in the same plane which is perpendicular to the mirror's surface. At the point of incidence on the mirror we establish the normal direction to the surface. The following is therefore valid:

Incident ray, reflected ray and normal all lie in one plane.

Angle of incidence = angle of reflection

This is called the reflection law.

Exercises

1. Fig. 21.28a shows, from above, two flat mirrors positioned at a right angle. Parallel light shines in from below, left. Sketch how the paths of the light beams A and B continue.

2. Fig. 21.28b shows a curved reflecting surface with light falling upon it from a uniform direction. In every point the direction normal to the surface is in the plane of the drawing. Sketch the further path of he beams A and B.

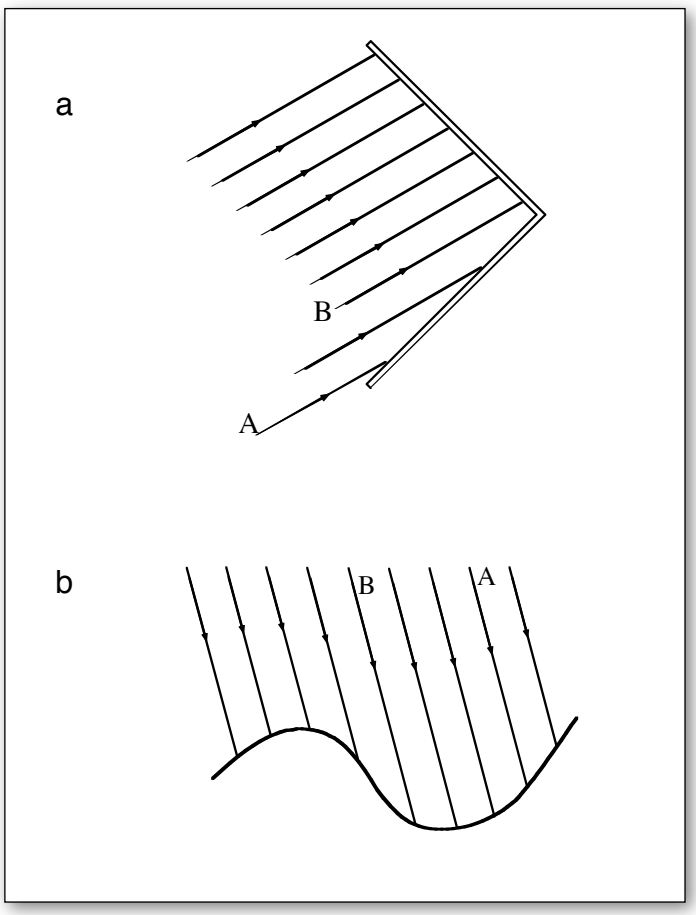

Fig. 21.28
For Exercises **1** and **2**

21.6 Plane mirrors

A bottle is standing in front of the plane mirror in Fig. 21.29. We see the bottle in front of the mirror and we see a second bottle that appears to be standing behind the mirror. The “phantom bottle” behind the mirror appears to be exactly as far away from it as the real one in front of it.

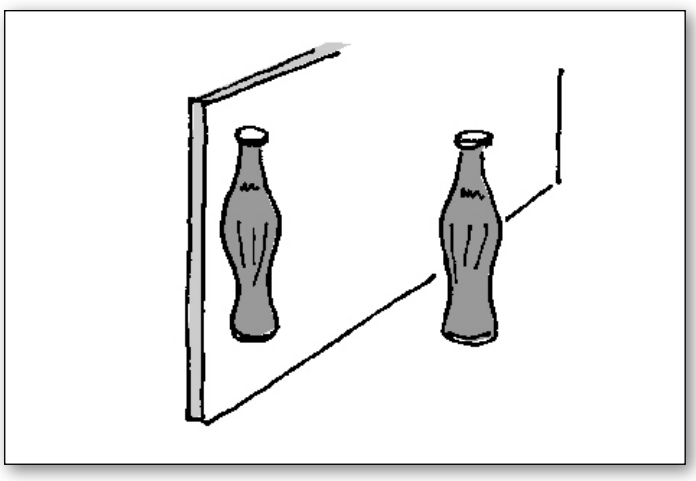

Fig. 21.29
A second bottle appears to be standing behind the mirror.

This is proven when a ruler is put there, Fig. 21.30. In addition, the false bottle stands exactly on the extended perpendicular line that can be drawn from the real bottle to the mirror. How does this “mirror image” come to be?

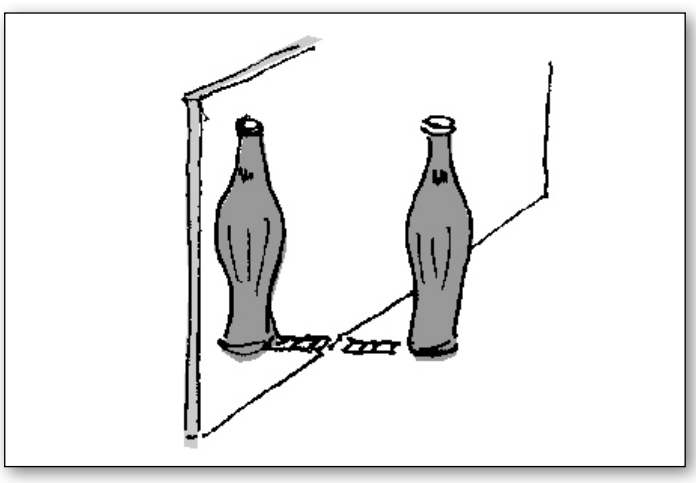

Fig. 21.30
The “fake” bottle stands at the same distance to the mirror as the “real” one.

In order to understand this, it is enough to apply the reflection law. P is a certain point on an object. Light is going away in all directions from P. In Fig. 21.31, three of the many rays going out of P are plotted. All three hit the mirror. The corresponding reflected rays are also drawn in. If the reflected rays are extended backwards—see the dashed lines—they meet at a point P'. The reflected rays come from the mirror, i.e., from points A, B, and C of the surface of the mirror. They appear to come from just one point P', behind the mirror, though.

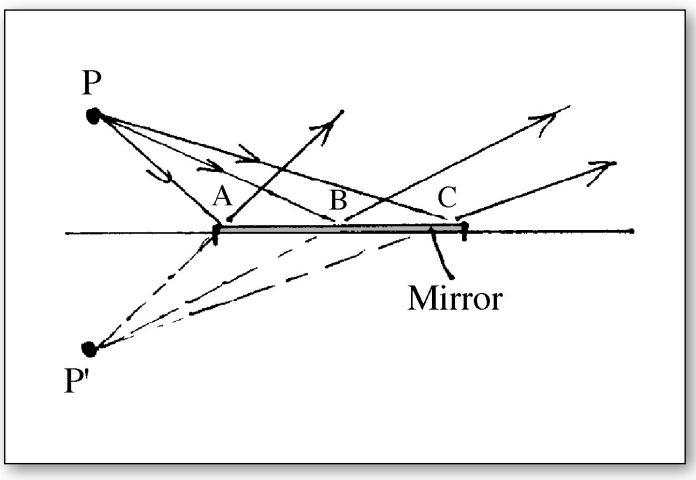

Fig. 21.31
The reflected light rays appear to be coming from point P’.

In Fig. 21.32, the mirror is replaced by an open window. In place of the light emitting point P' in Fig. 21.31, there is a real light emitting point in Fig. 21.32. The light coming out of the window in Fig. 21.32 cannot be distinguished from the light coming from the mirror in Fig. 21.31.

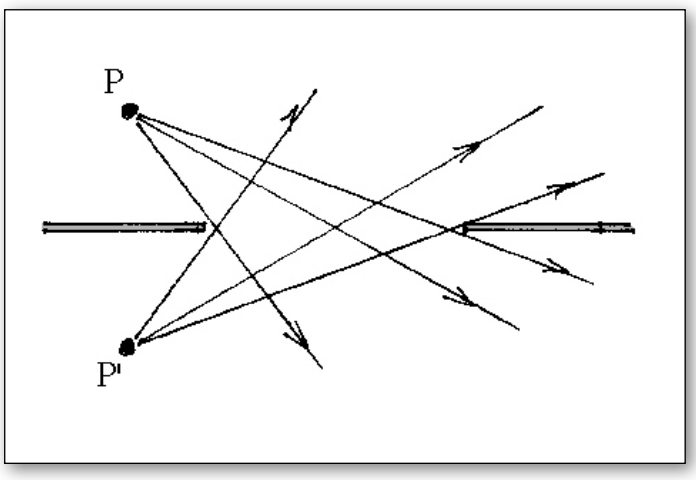

Fig. 21.32
The distribution of light above the window is the same as above the mirror in Fig. 21.31.

Exercise

Fig. 21.33 shows a mirror and a rod shaped object from above. Determine the position of the feigned object behind the mirror. Draw in some rays.

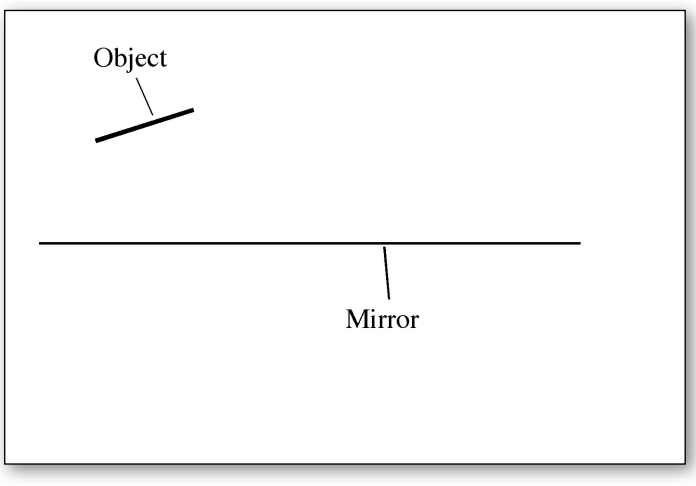

Fig. 21.33
For the exercise

21.7 Parabolic mirrors

We often need a lot of light at just one point. Light that comes in widely distributed should be concentrated onto a small spatial area. We will consider an example of how this is done.

The boiler of a power plant is to be heated with sunlight. If the boiler were just put into the sun, it would hardly heat up at all because not enough sunlight shines upon it. The sunlight that would ordinarily shine upon a larger area must be collected somehow. This is done with the help of mirrors, Fig. 21.34.

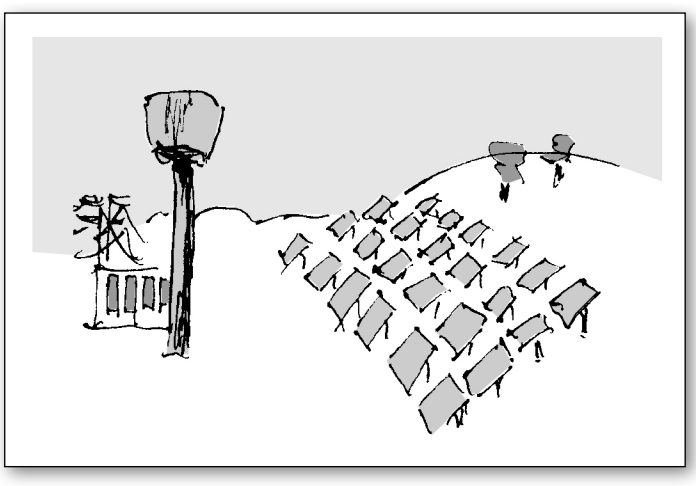

Fig. 21.34
Solar power plant: The mirrors concentrate the sunlight upon the boiler at the top of the tower.

Each mirror is oriented so that the sunlight reflected from it falls upon the boiler. In order for the light from the greatest possible surface to be used, the boiler is installed in a tower. This is why it is called a solar tower power plant.

It is basically possible to use just one large mirror instead of many individual ones, Fig. 21.35. A power plant would need such a huge mirror, though, that it would be completely impractical. It would have no advantages over the many small ones. The method is practical, though, for smaller systems.

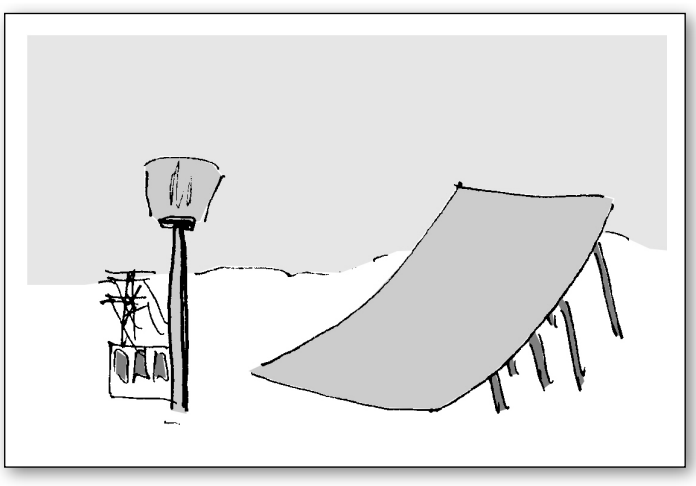

Fig. 21.35
It would be possible to replace all the individual mirrors in Fig. 21.34 with one large curved mirror.

If light shines from one direction, if it is parallel light, it can be concentrated with a mirror onto a single point. The mirror must have the right form to do this though. The cross section of its surface must be a parabola. This kind of mirror is called a *parabolic mirror*.

Light that falls parallel to the symmetry axis of a parabola is reflected to meet at one point. This point is called the *focal point* of the mirror.

Sunrays are not quite parallel, so sunlight is not concentrated into a point, but into a small spot.

A parabolic mirror can concentrate parallel light onto one point. It can do just the opposite as well. It can make the light coming from one point into parallel light.

We take a light source that is “as point-like as possible”, a light bulb with a very compact filament, for example. We position the lamp so that the filament is as precisely as possible at the focal point of the parabolic mirror. The mirror now reflects the light hitting it so that it is (almost) parallel. We have built a spotlight similar to a car’s headlight or a flashlight.

Parabolic mirrors have many more uses. They can be used in both ways in order to concentrate parallel light onto a point or to make the light coming from a very small, point-like source parallel. They are often used for light that is far beyond visual sensitivity of our eyes. These mirrors are used for sending and receiving antennas of various kinds of electromagnetic waves.

The parabolic mirror of a sending antenna creates a beam of relatively parallel “light” out of the actual rays of the almost point-like antenna. The mirror of the receiving antenna then collects the rays falling upon it and concentrates them upon the actual small receiving antenna.

Such sending and receiving antennas can be found on telecommunications towers. With their help, television programs, radio programs, and telephone calls are transmitted from one tower to the next.

Parabolic antennas are also used to send and receive data from satellites. An antenna with a parabolic mirror is used for direct reception of satellite television.

The parabolic mirror of the radar at an airport serves as sender and receiver simultaneously, Fig. 21.36. It creates a relatively thin beam of electromagnetic waves. The antenna revolves so the beam moves in a circle. The “light” it sends is only reflected by metal. If the ray hits an airplane, it is reflected, concentrated by the parabolic mirror, and then received by the actual antenna at the focal point. This is how the direction of an airplane is determined.

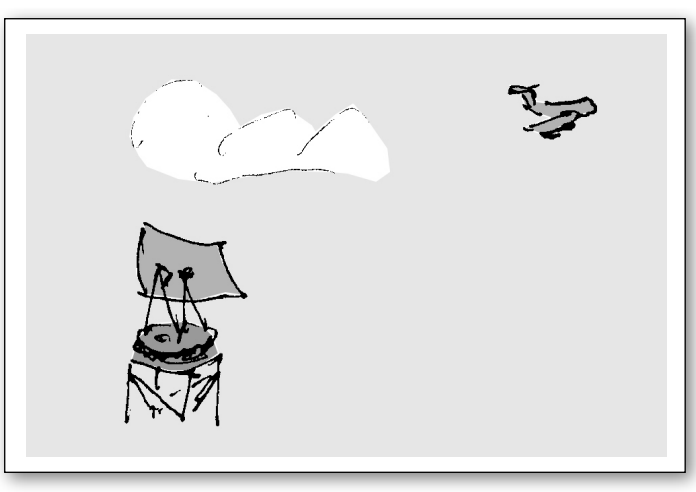

Fig. 21.36
Radar antenna

Now, back to the “visible” light.

What does the field of mirrors in Fig. 21.34 do when the weather is overcast? We will consider just one mirror. It reflects the light falling upon it from a single direction onto the boiler. Most of the light falling upon it is deflected in other directions and is lost. No matter how the mirror is turned, most of the light goes off-target.

This is also true for parabolic mirrors because no mirror can concentrate diffuse light.

Diffuse light cannot be concentrated.

Exercise

The parabolic mirror in Fig. 21.37 has diffuse light falling upon it. Show that the mirror cannot concentrate it and make it parallel.

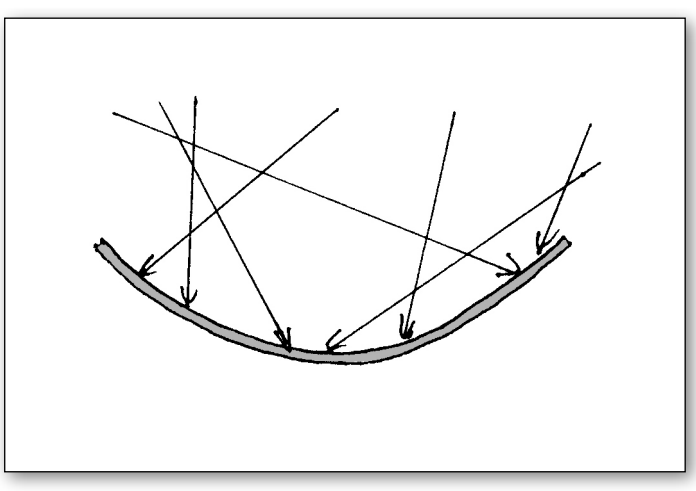

Fig. 21.37
For the exercise

21.8 Refraction of light

We need a thin ray of light for the following investigations.

We fill an aquarium with water and put some drops of milk in it. The water becomes a bit cloudy. If the light ray now shines through the water, it is easy to see what path it takes.

We send the ray of light into the water from above, Fig. 21.38. The most important observation: The ray is bent where it enters the water except when it shines perpendicularly onto the water's surface. The greater the angle between the ray and the direction perpendicular to the surface, the more it is bent (away from the surface).

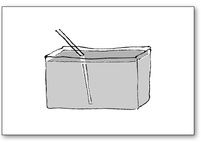

Fig. 21.38
The light beam is bent at the water's surface.

This bending of light rays is called *refraction*. One says that light is refracted.

Fig. 21.39 shows incident and refracted rays. The normal at the point of impact is also drawn in. The angle of the ray to the normal is labeled α, and the angle between the refracted light ray and the normal is labeled β.

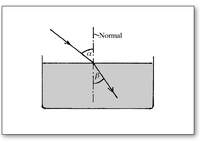

Fig. 21.39
When it hits the water's surface, the light is refracted toward the normal.

We can therefore say that the light is refracted toward the normal at the transition from air to water. We also find that incident light and refracted light lie at a plane that is perpendicular to the refracting surface. In other words: Incident ray, refracted ray, and normal all lie in the same plane.

When the light ray enters other transparent materials from air, it is refracted as well. How strongly it is refracted depends upon the material. Light is more strongly refracted at the transition to glass than at the transition to water, and more strongly with diamonds than with glass.

Table 21.1 shows the relation between α and β for water, glass and diamond.

Table 21.1
The relation between angles α and β for water, glass, and diamond

	β		
α	water	glass	diamond
0°	0°	0°	0°
10°	7,5°	6,6°	4,1°
20°	14,9°	13,2°	8,1°
30°	22,1°	19,5°	11,9°
40°	28,9°	25,4°	15,4°
50°	35,2°	30,7°	18,5°
60°	40,6°	35,3°	21,0°
70°	45,0°	38,8°	22,8°
80°	47,8°	41,0°	24,0°
90°	48,8°	41,8°	24,4°

There are still some simple questions to be discussed.

What happens with light that passes from water into the air? Instead of putting the light source in the water tank—that would not do it much good—we will use a trick. We send the light ray into the water from outside and put a mirror into the water so that the light ray falls upon it perpendicularly. The light ray is reflected into itself in the water. What happens with the reflected ray at the surface of the water? The experiment shows clearly that it is again refracted in exactly the same direction from which it came, Fig. 21.40.

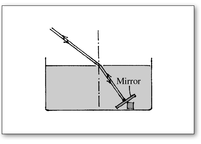

Fig. 21.40
When the light leaves the water, it is bent back to where it came from.

When the light enters the water it is refracted toward the normal, and when it leaves the water, it is refracted away from it.

The same applies for other transparent materials. We can now imagine what happens when a light ray traverses a glass plate, Fig. 21.41. At the point where it enters the glass plate, it is refracted toward the normal, and at the point of exit, it is refracted away from it. In all, it simply undergoes a parallel shift.

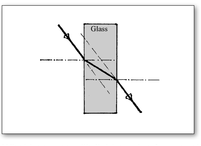

Fig. 21.41
When passing through a glass plate, the light ray undergoes a parallel shift.

What happens with light coming from another material than air when it enters glass? There is one special case that allows an easy answer, the case of light moving from glass into glass, Fig. 21.42. When it leaves the glass block on the left, it is refracted away from the normal. When it enters the block on the right, it is immediately refracted back into its original direction. All in all, nothing happens to it.

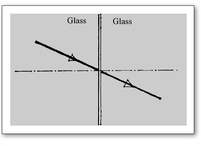

Fig. 21.42
When passing from glass to glass, the light does not change direction at all.

So far we know that:

The transition from air to glass means strong refraction and from glass to glass means no refraction.

The transition from water to glass lies between these two cases. The light is not as strongly refracted as when it comes from the air, but it is more strongly refracted than when it comes from glass. We therefore know:

The transition from water to glass means weak refraction.

We say that the different transparent materials have different *optical densities*. Of the three substances in Table 21.1, diamond is the optically densest, followed by glass and then water. Air has a smaller optical density than the three substances just mentioned. A vacuum has an even smaller one than air. The difference between air and vacuum is very small, though.

We now have the rule:

> When light moves from material A to material B, and if the optical density of B is greater than that of A, it is refracted toward the normal. If the optical density of B is smaller than that of A, the light is refracted away from the normal.

The optical density of air and all other gases depends upon their densities (mass per volume). The density of air can be changed easily by heating the air. The experiment in Fig. 21.43 makes use of this effect.

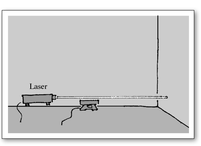

Fig. 21.43
The hot air above the hotplate is less dense than the air surrounding it. Thus, also its optical density is lower.

A laser beam is sent very closely over the surface of a heated hotplate and falls somewhere upon the wall. If one lightly blows the air above the hotplate, the spot of light on the wall also moves (be careful, never look into the beam!). It looks like the laser beam is being blown away. Actually, only the hot air is being blown away. As a result, the transitions between hot and cold air that caused a slight refraction of the laser light, disappear.

Varying optical density and the refraction of light related to it is also the cause of objects seeming to flicker when they are observed close above a heater.

Exercises

1. Represent the relation between α and β for water, glass, and diamond in an α-β coordinate system.

2. Fig. 21.44a shows a light ray entering glass from the air. Sketch the refracted light beam.

In Fig. 21.44b, a light ray again enters glass from the air. Only the refracted ray is plotted. Sketch the ray falling upon the glass.

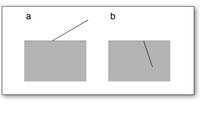

Fig. 21.44
For exercise **2**

21.9 Prisms

In the last section we saw that a light beam falling upon a glass plate with parallel surfaces will undergo a parallel shift to the side, Fig. 21.41. We will now change the situation a little and have the light beam pass through a glass body defined by flat surfaces that are not parallel to each other. The simplest body of this type is a prism with a triangular base, Fig. 21.45.

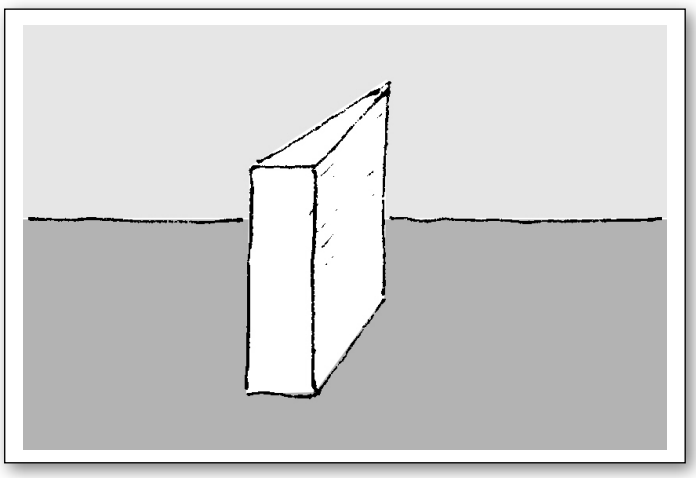

Fig. 21.45
Prism with a triangular base

The beam is refracted twice, where it enters the prism, and where it leaves it. All the beams and normals lie in one plane that is parallel to the base of the prism. This plane is shown in Fig. 21.46.

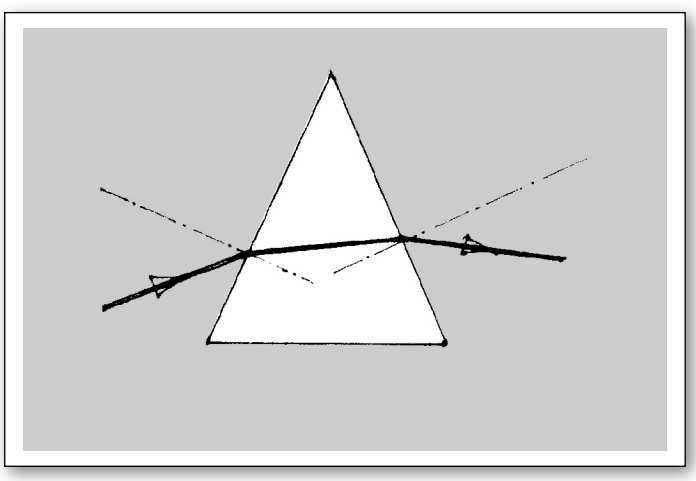

Fig. 21.46
After two-fold refraction of the light ray, there is a net deflection.

We see that the light leaving the prism has a different direction than the light entering it. This is different than with the glass plate. It undergoes a net deflection.

The total deflection depends upon the angle at which the light hits the prism.

Earlier we used the fact that a prism deflects light of different wavelengths (different types of light) differently.

Exercises

1. A glass prism has a base with three equal sides, Fig. 21.47. Light having a uniform direction falls upon the prism from the left side. Determine the direction of the light exiting the prism.

2. Two identical prisms are put one behind the other, Fig. 21.48. What is the light's direction after it has passed through both prisms?

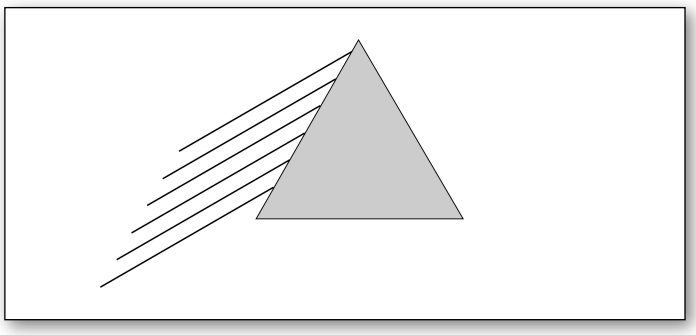

Fig. 21.47
For exercise **1**

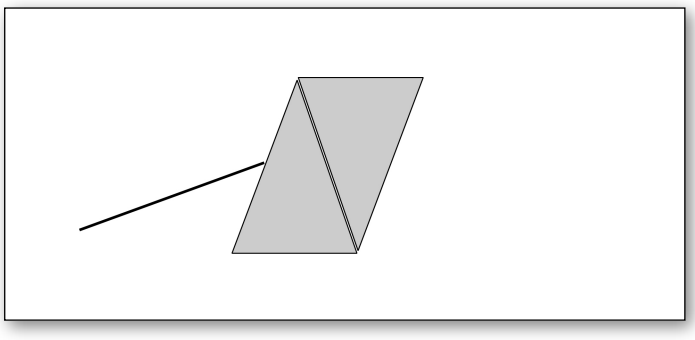

Fig. 21.48
For exercise **2**

21.10 Total reflection

Table 21.1 shows that a light ray entering water at an angle α of almost 90°, continues underwater at an angle of 48.8° towards the perpendicular. These numbers also mean that a light ray falling upon the water's surface at 48.8° from below in the water, will run almost parallel to the surface. What happens with a light ray that hits the surface of the water from below, but at a greater angle? It cannot exit the water at all. It is reflected back down into the water exactly as stated by the reflection law, Fig. 21.49. This phenomenon is called *total reflection*.

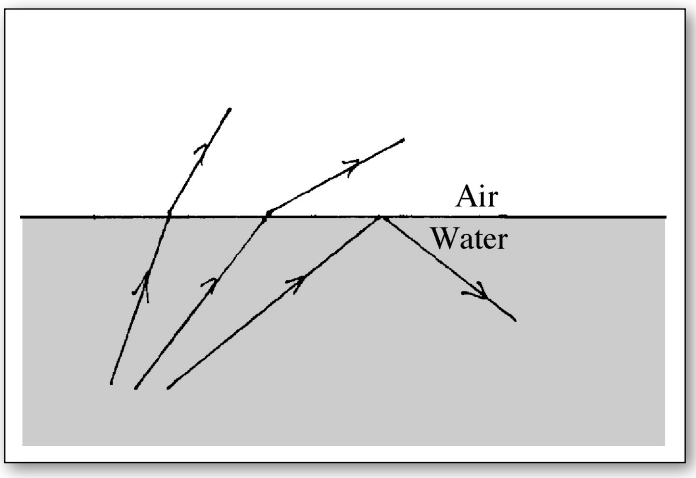

Fig. 21.49
Light coming from below the surface of water will be totally reflected by the surface if the angle of incidence is greater than 48.8°.

This reflection doesn't suddenly happen when the angle β reaches 48.8°. In general, only a part of the light is refracted at the boundary layer; the rest is reflected. Moreover, the steeper the direction of the incident light, i.e., the smaller the angle to the normal is, the less is reflected.

Optical fiber cables are an important application for total reflection. An optical fiber is a long flexible glass fiber. Light entering it at one end, at a small angle to the perpendicular, cannot leave the optical cable sideways; it is totally reflected, Fig. 21.50. It runs in a zigzag course through the cable and follows its bends. It exits the cable at the other end.

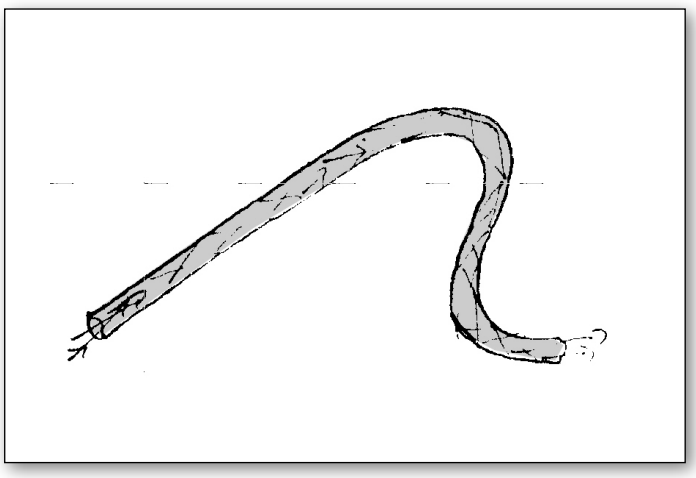

Fig. 21.50
Optical fiber

The remarkable thing about these optical fibers is that it has become possible to produce types of glass that can allow light to run several hundred meters without much loss due to absorption. Consider that the light that falls upon the surface of the ocean is practically nonexistent at a depth of 300 m. It is pitch black at that depth—even when the water is very clean.

Exercises

1. How does the light in Fig. 21.51 continue? Take into account that sometimes a part of the light is refracted and a part is reflected.

2. A ray of light falls upon a cylindrical glass rod, Fig. 21.52. Sketch how the ray continues.

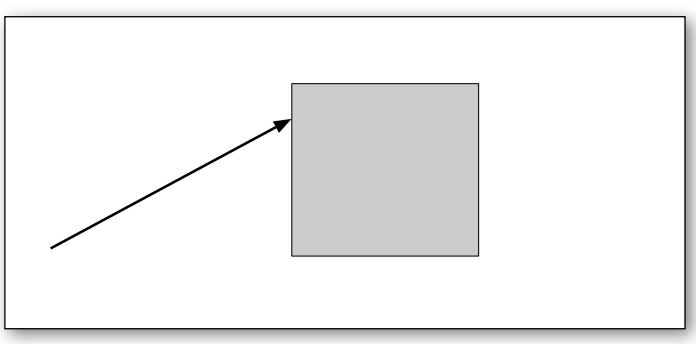

Fig. 21.51
For exercise **1**

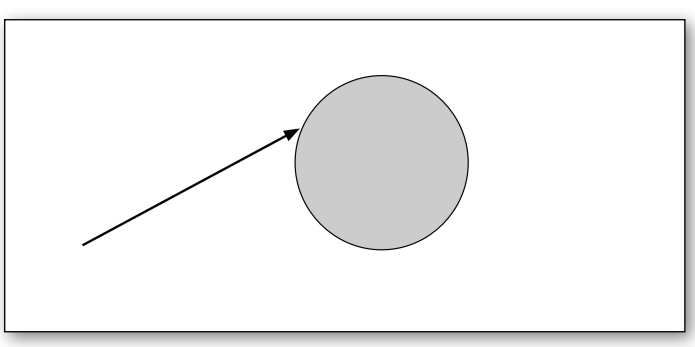

Fig. 21.52
For exercise **2**

22

Optical Image Formation

22.1 What is an image?

We have already encountered light as an energy carrier. Light carries energy from the Sun to the Earth. Without this light, it would be so cold on Earth that life would not be possible.

We have also seen that light can be a data carrier both in nature and technology. With light, people and animals receive large amounts of data through their eyes. They use these data to find their way in the world. Light is used as a data carrier:

– to transport large amounts of data in optical fibers;
– to read the music on a CD or the images an a DVD;
– to read the sound track of a film.

There is another area of technology that deals with light as a data carrier and has developed into a specialized field of its own. This field is called *optics*. The most important purpose of optics is producing so-called *optical images*. Before we begin our investigation of optical image formation, we will answer the question of what an image actually is. The answer is not as simple as you might think.

Is a picture an object, a thing? It seems so. A painting, a photograph, a drawing – these are all pictures. These pictures remain pictures even when they are put into the dark where they can no longer be seen. A passport picture is still a passport picture even when the passport is in a pocket.

We will now consider a picture that is produced upon a wall when a slide or a movie is projected upon it. If the projector light is turned off, the picture disappears from the wall. No one would even consider saying that the picture has remained on the wall.

Despite this difference, a fleeting projected picture has a basic similarity to a "material" picture, say a paper photograph, provided that the photograph is being lit. It also has something in common with another kind of fleeting image: The picture on a television screen.

We will compare three types of images: 1. an image on paper; 2. a projected image; 3. an image on a television screen.

In order to make the problem clear, we will assume that the imaged "object" is made up of luminescent points in dark surroundings. The luminescent points are all on one plane. This plane is shown from the side in Fig. 22.1.

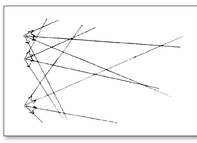

Fig. 22.1
Light distribution in front of an object consisting of three luminous points

First we consider the light being emitted by the "object" itself. Light rays emanate from each point and go off in all directions. No light comes from anywhere else.

Now we look at an image on paper that shows these luminescent points, Fig.22.2. The light falling upon the paper is absorbed everywhere except where we have images of the bright points. From there it is reflected and scattered. The light emitted by the paper image is distributed exactly like the light being emitted by the actual luminescent points in Fig. 22.1.

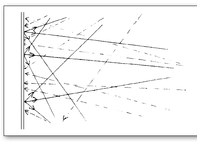

Fig. 22.2
Distribution of light in front of a paper image of the "object" in Fig. 22.1. The light falling onto it is represented by dashed lines.

You see that it is important that the light being thrown back is scattered as well. If it were being reflected only, the bright points would be visible from only one direction.

Let us now turn to the projected image in Fig. 22.3. The screen itself does not absorb light. It reflects and scatters all the light that falls upon it. However, light only falls upon a few places on the screen, and it is back-scattered from these places only. The distribution of the light coming off the screen is, again, the same as that of the original object in Fig. 22.1.

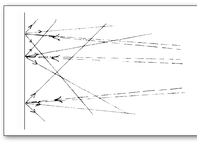

Fig. 22.3
Distribution of the light in front of a projected image of the "object" in Fig. 22.1. The light falling onto the screen is represented by dashed lines.

The difference between a picture on paper and a projected picture can be described as follows. The data on a paper picture are stored there. They stay there even when the paper is in the dark. The data for projected slide or film images are stored on the slide or film. This data are read out while the slide is being projected. The same data are read out again and again because the picture is still.

We will now look at an image on a television screen. In contrast to a picture on paper or a projected image, a television screen is not illuminated from outside. The light comes from the screen itself. Again, we will assume that the "object" being imaged is the same one as in Fig. 22.1. This time, light rays are being emitted in all directions from points on the screen, Fig. 22.4. The distribution of light is, as before, the same as that of the original object in Fig. 22.1.

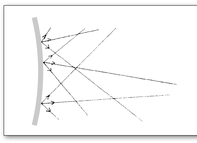

Fig. 22.4
Light distribution in front of a television image of the "object" in Fig. 22.1

A characteristic of images is that they are flat or "two-dimensional". They have no physical depth.

Our simple object of points lying next to each other was two-dimensional as well. Now, how would the light distribution look for a three dimensional object, one with depth?

You surely already know how a three-dimensional house is drawn upon a two-dimensional piece of paper. The third dimension is flattened out.

Again, we will consider a simplified model. Fig. 22.5 shows the light distribution of the original as seen from the side. In this case, the points of light are not on just one plane. Here, points A, B, and C lie on plane 1, and point D lies on plane 2.

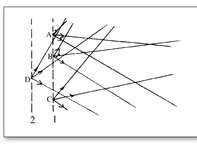

Fig. 22.5
The object points are not on a single plane anymore.

Now we will investigate an image of these points, possibly a photograph, Fig.22.6. In the process of making the photograph, the points are all pushed together onto one plane. We see that the distribution of light coming from the picture is not the same as that which comes from the original. It is different in three-dimensional reality than in a two-dimensional image.

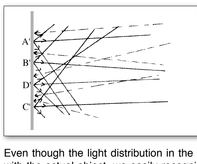

Fig. 22.6
An image of the "object" in Fig. 22.5. The luminous points are pushed together onto one plane.

Even though the light distribution in the picture is different than it is with the actual object, we easily recognize the object in the picture. On the other hand, we see clearly that it is only an image. We would never mistake it for the actual object.

There are images that can exactly reproduce the light distribution of a three dimensional object, Fig. 22.7. They are called holograms. In order to produce the correct light distribution, the illumination must be done with light from one direction only. Laser light is best suited for this.

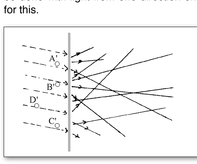

Fig. 22.7
A hologram produces the same light distribution as the original object it depicts.

Exercises

1. You probably know that it can be difficult to make out a picture on a slide when you hold it in your hand. Why is this? How should you hold the slide in order to see clearly what is on it?

2. Would it be better to project a slide onto a mirror? What would we see?

22.2 Pinhole cameras

A box with a small hole in the middle of one side is the simplest device for making an optical image. On the side of the box opposite the hole is a ground glass or parchment paper, a material that lets light through and diffuses it. An image of objects outside the box on the side with the hole can be seen on the screen. How does this happen?

For the sake of clarity, we will represent the “landscape” we wish to show as three luminous points all lying on one plane. Fig. 22.8a shows this plane from the side as well as the points and the light being emitted by them.

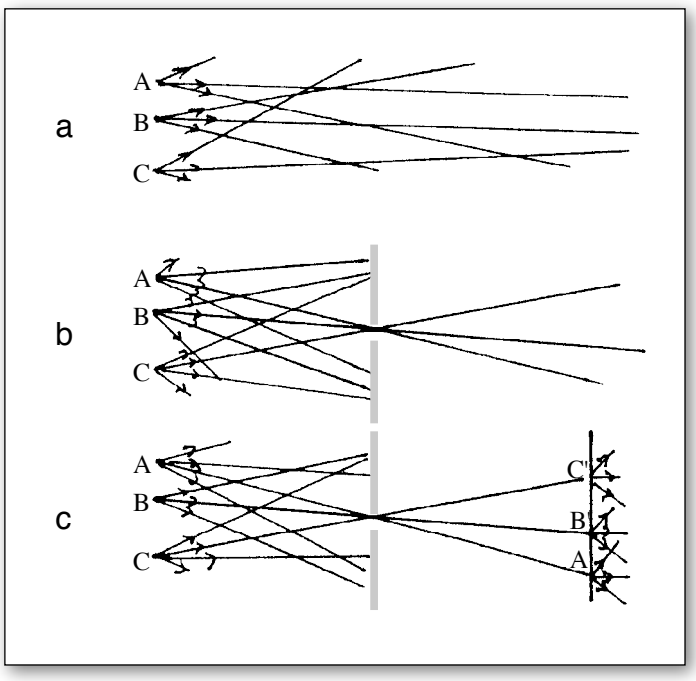

Fig. 22.8
The light coming from the object first meets an aperture and then a ground glass screen.

First we will block this light with an aperture plate, Fig. 22.8b. Only three thin beams of the light going to the right are allowed through the aperture.

We allow these light beams to fall upon a ground glass, Fig. 22.8c. The light coming through the pinhole hits the screen in three places and is scattered. We now have three luminous points A', B' and C' on the ground glass. A' is an image of A, B' is an image of B und C' is an image of C. The distribution of light is the same as that of the three luminous points A, B, and C, only their order has been reversed. We can say that the image stands on its head.

In our example, the luminous “object points” were lying on a single plane, and the ground glass screen of the pinhole camera was parallel to this surface. The object was as flat as the screen. Now we will assume that the object has “depth”: One of the points lies further back than the others, Fig. 22.9. The screen is flat so we obtain a flat picture. Our “landscape” has, again, been flattened.

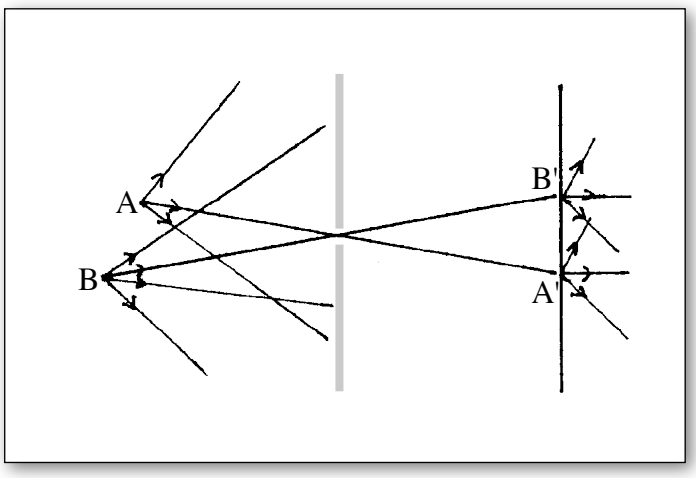

Fig. 22.9
The points of the object no longer lie upon one plane.

An especially impressive variation of the pinhole camera is the *camera obscura:* A darkened room with a small opening at the window. On the white wall opposite this small opening, a picture of the landscape outside the room can be seen, Fig. 22.10.

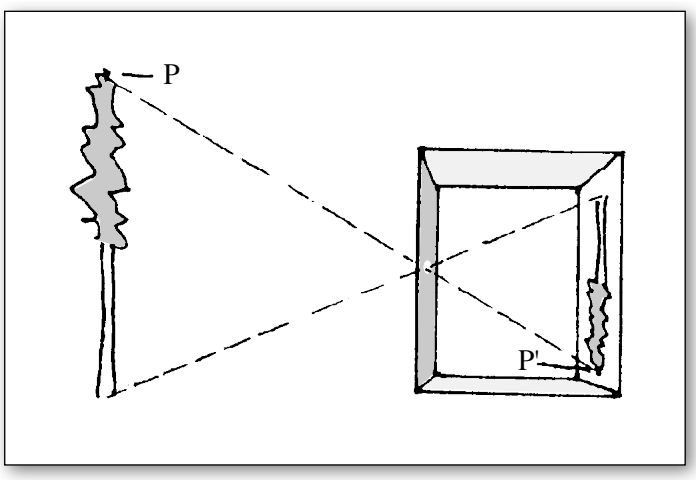

Fig. 22.10
A camera obscura. The image is seen from inside of the “pinhole camera”.

The hole again ensures that the light falling anywhere upon the wall comes from only one direction. The light from the top of the tree P falls only upon the location P’ on the wall. The white wall is needed for scattering the light.

Exercises

1. Why is there no image to see if a transparent glass plate is used instead of a ground glass screen in a pin hole camera?

2. What would we see on the screen of a camera obscura if the screen were a mirror? The object being pictured is a single luminous point. What would the light distribution be?

22.3 The relation between object size and picture size

Fig. 22.11 shows a schematic of the image of an object with a pinhole camera.

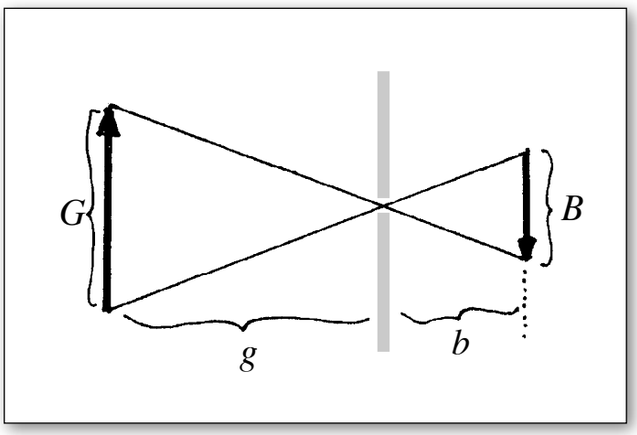

Fig. 22.11
G is the size of the object, *B* is the size of the image, *g* is the distance of the object and *b* is the distance of the image.

We will call the size of the object *G* and the size of the picture of it *B*. The distance *g* of the object from the pinhole is called the object distance, the distance *b* of the picture from the aperture is called the image distance. There is a simple relationship between the four quantities *G*, *B, g* and *b*:

$$\frac{G}{B} = \frac{g}{b}$$

If you already know the theorem of intercepting lines from your math class, you can read the equation directly from Fig. 22.11 right away. If you don't know this theorem, you can convince yourself of the validity of the equation by testing it.

For example, in Fig. 22.11

$G = 30$ mm

$B = 15$ mm.

This results in $G/B = 2$.

Moreover, we measure

$g = 40$ mm

$b = 20$ mm.

This results in $g/b = 2$. The equation $G/B = g/b$ is satisfied.

Exercises

1. A pinhole camera takes a picture of a church tower. The church tower is 100 m away from the camera. The distance between the pinhole and the screen is 16 cm. The picture of the church tower is 8 cm high. How high is the actual church tower?

2. The cathedral in Cologne is 157 m high. By using a 20 cm long pinhole camera, you produce a picture in which the towers are 2 cm high. How far away is the cathedral?

3. A tree with a height of 8 meters produces a 1 m high picture in a camera obscura. How high is a second tree that stands right next to the first one and whose image is 1/2 m high.

22.4 Improving the pinhole camera

We construct a camera obscura. On the wall we see a picture of the house across the street. However, the picture is very dark. Maybe you think that this is easy to fix by just making the hole a little bigger. We try it out and yes – the picture becomes brighter. Something else happens as well, though, something we don't want. The picture becomes blurred. You can see why this happens by looking at Fig. 22.12.

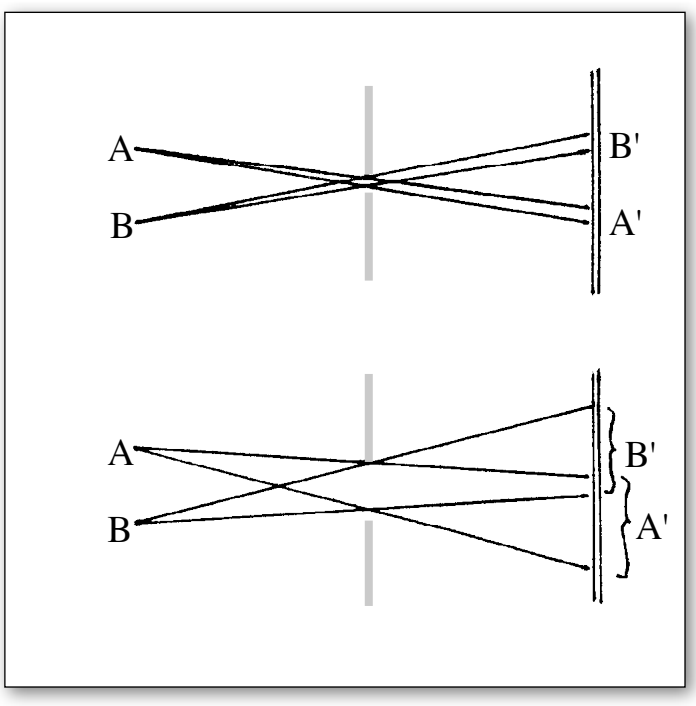

Fig. 22.12
Small hole: The images are separate from each other. Large hole: The images overlap.

In the upper picture, the hole is small. The image of the two object points A and B are two tiny luminous points A' and B'. In the lower picture, the hole has been enlarged. The "pictures" of A and B are now two spots that overlap. It is difficult to distinguish A' from B'. The picture of points A and B is *blurred*.

We can make a general conclusion that:

> The larger the hole of a pinhole camera, the brighter and more blurred the picture will be.

Simply enlarging the hole does not solve our problem. We need a better idea.

We begin again with a small hole and obtain a sharp but dark image. Now we make a second small hole at some distance from the first one. What can be seen on the wall? We get a second picture, Fig. 22.13a. Or better, the same picture twice. They are shifted relative to each other.

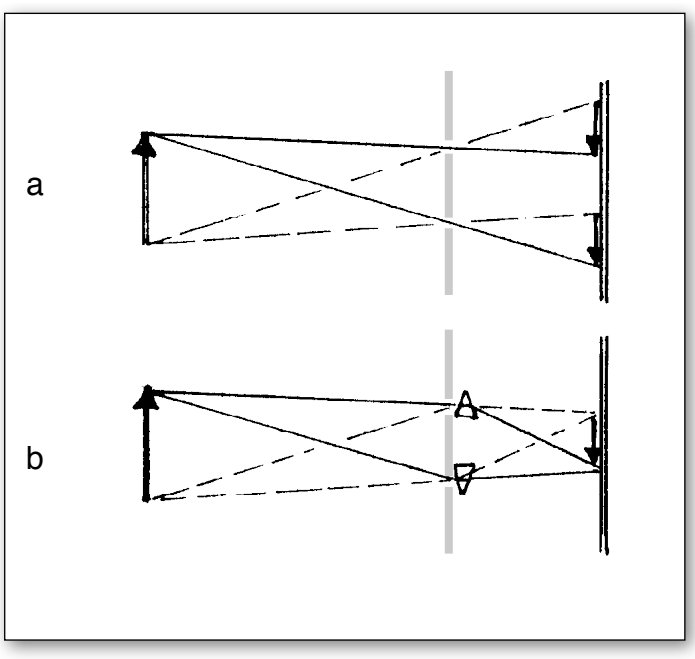

Fig. 22.13
(a) The pinhole camera with two holes produces two images.
(b) The two images are brought to coincidence by prisms.

We are a little closer to the solution of our problem now. We only need to move the pictures so that they lie exactly on top of each other. We have to bend the light beams producing the two images, into the middle. You know how to do this: we can use prisms, Fig. 22.13b. The result is a picture that is twice as bright as the one from only one hole and it is sharp!

There is one disadvantage in all this, however. Our first pinhole camera – the one with one hole – always gave us a sharp image no matter how far the hole was from the screen. This distance only affected the size of the picture, Fig. 22.14a.

Fig. 22.14
(a) In a single-hole camera, the size of the image is the only thing affected by the distance of the aperture to the screen.
(b) To produce a sharp picture with a two-hole camera, the distance of the screen to the aperture must be set to a particular value.
(c) No prism needs to be put in front of the hole in the middle.

This is different with double pinhole cameras. The two images in Fig. 22.13b only match up if the screen is at a certain distance from the hole. Three screen positions are represented in Fig. 22.14b. The images only match up when the screen is in position 2. If it is more to the front, in position 1, or further back, in position 3, they do not match up.

In Fig. 22.14c, a third hole is made in between the first two. We do not need to put a prism in front of this one. The corresponding image automatically matches up with the other two. We find that, again,

$$\frac{G}{B} = \frac{g}{b}$$

holds for the light beam running through the middle hole.

Exercise

Mirrors can be used instead of prisms to bend light. How would a "two-hole-camera" look if mirrors were used in it instead of prisms? How would a corresponding camera with lots of prisms look? What kind of mirror would we have if all the individual mirrors were assembled into one continuous mirror?

22.5 Lenses

If we wish to make our picture even brighter, we know how to do it. We simply make more holes and put the appropriate prism in front of each one.

We can ultimately make one big hole, like we tried before. However, we have to cover the entire cross sectional area of the hole with prisms, Fig. 22.15.

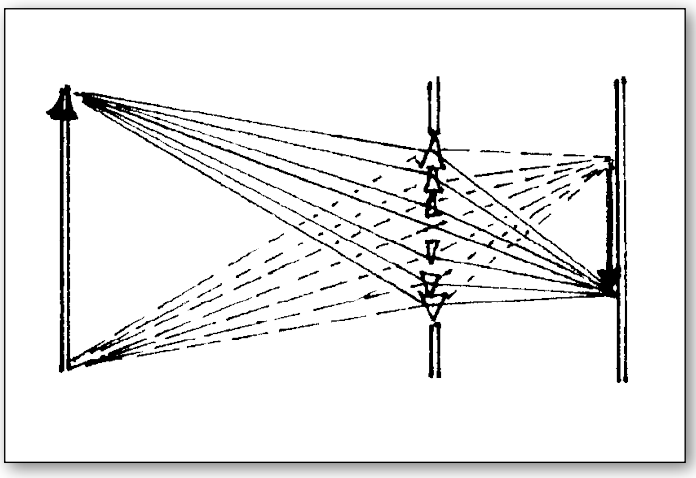

Fig. 22.15
The entire area of the large opening is covered by prisms.

The light passing through the middle of the hole does not need to be refracted at all. The further away from the middle the light coming through is, the more strongly it must be bent toward the middle. The prisms gradually become flatter toward the center. Fig. 22.16a shows how the prisms look from the side. Now, the hole does not have only a vertical direction, but a lateral one as well. The prisms are therefore gradually flatter in form as they go from the rim toward the middle.

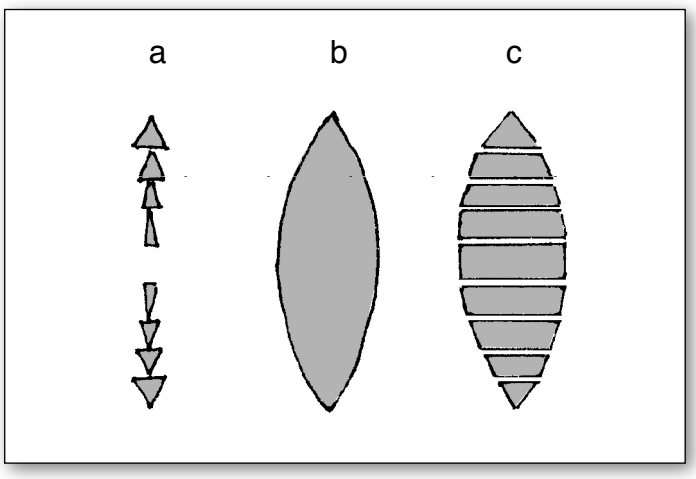

Fig. 22.16
Instead of using single prisms (a) a single glass body can be used (b). This can be thought of as being made up of many prisms (c).

Instead of using many individual prisms, a single piece of glass can be used, i.e., a *lens*, Fig. 22.16b. You can imagine the lens as being made up of many small prisms, Fig. 22.16c. The prisms in 22.16c are thicker than those in 22.16a, but as only the angle between the surfaces opposite each other counts, the optical image formation can be achieved with the thicker prisms.

A lens can have a different form than the one in Fig. 22.16b and still have the same effect, Fig. 22.17. The only important thing here is that the angle of the lens surfaces opposite each other increases from the middle outwards.

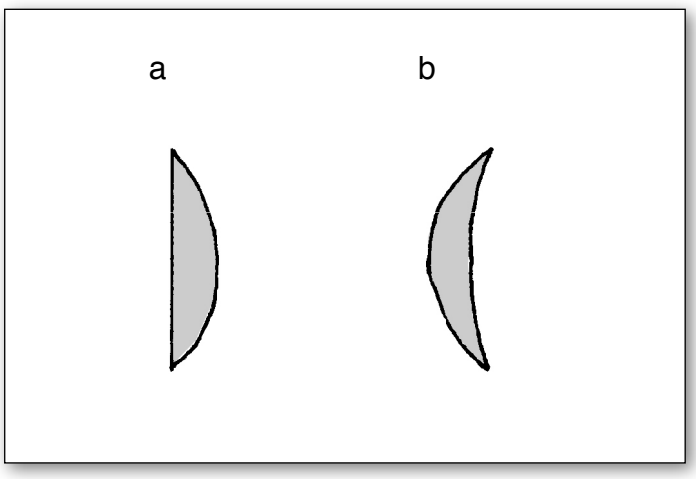

Fig. 22.17
Two different lens forms

The two surfaces of most lenses are curved. Sometimes one side is flat, Fig. 22.17a.

Fig. 22.18 shows how rays run when an object is imaged by a lens.

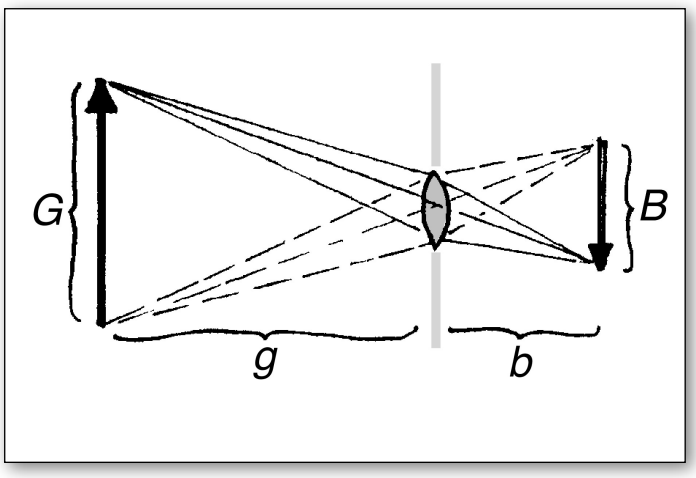

Fig. 22.18
An image of an object is made by a lens.

The optical imaging process of a lens is actually the same as the one using the improved pinhole camera we investigated in the last section, so:

$$\frac{G}{B} = \frac{g}{b}$$

holds for it as well.

22.6 Making optical images with lenses

Our improved pinhole camera showed us that a picture is only sharp (focused) when the screen is at the correct distance from the aperture. Correspondingly, it must hold that when we use a lens, the picture is only focused if the screen is at the correct distance from it. We will try it out, Fig. 22.19.

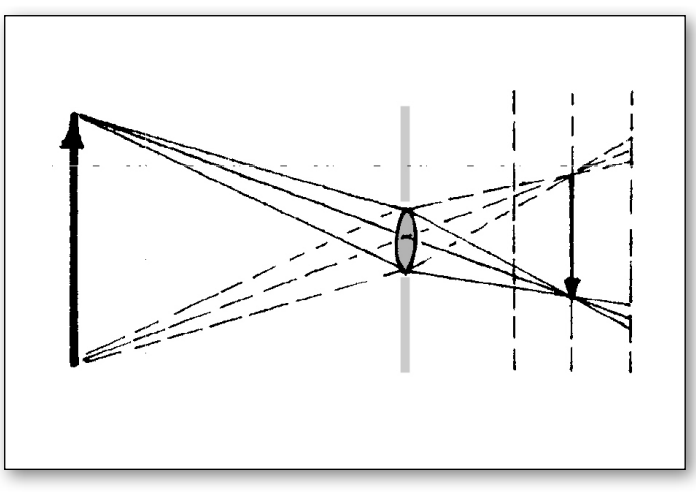

Fig. 22.19
The image is sharp only when the screen is at a certain distance (at fixed object distance).

The straight line passing through the middle of the lens is called the *optical axis*. We place the screen perpendicular to the optical axis and move it back and forth parallel to it. We find that the picture is only sharp when it is at a certain distance *b* from the lens, i.e., when the image distance *b* has a specific value.

One says that the image is situated at the distance *b* from the lens.

We leave the screen where we have the focused image and now move the object back and forth parallel to the optical axis, Fig. 22.20. This means that we change the object distance *g*. The picture becomes blurred when we move the object toward the lens, Fig. 22.20b. It is also blurred when the object is moved away from it, Fig. 22.20c. If the screen is not moved, the picture is only sharp when the object is at a certain distance from the lens.

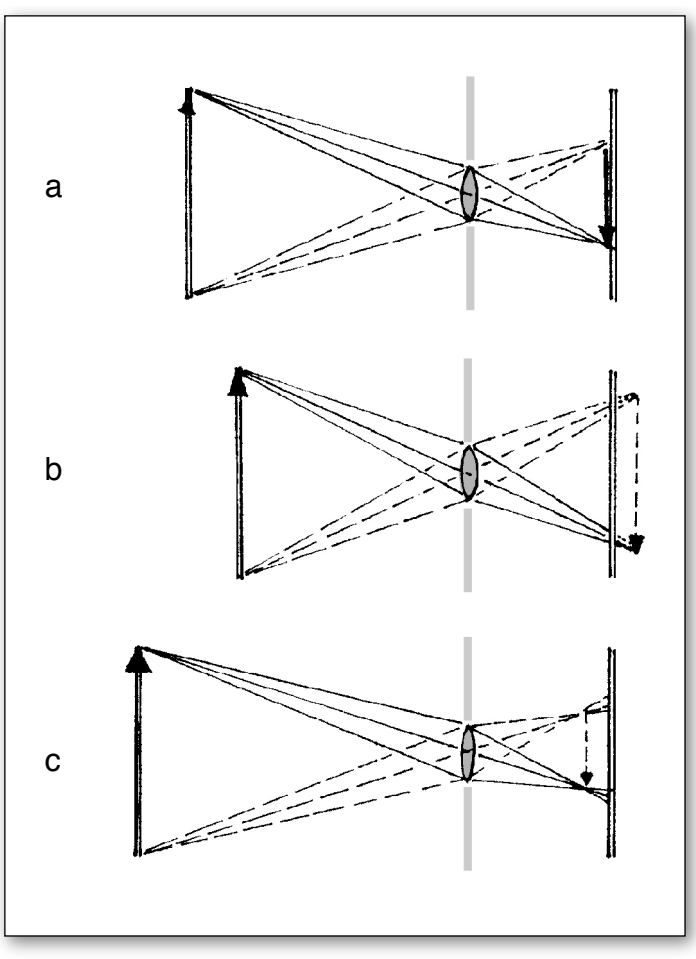

Fig. 22.20
(a) Initial setup.
(b) The object distance is decreased.
(c) The object distance is increased.

Let us do a third experiment. We again start where the positions of the screen and the object produce a sharp image. We move the object a little toward the lens and the picture becomes blurred. Now we can make the picture sharp again by either putting the object back where it was, or by moving the screen away from the lens.

We move the object once again, this time away from the lens. By moving the screen toward the lens, we focus the picture again.

Our observations allow us to make the following statement:

To every object distance *g* corresponds a certain image distance *b*. The greater *g* is, the smaller *b* must be.

Object and image distances are measured on the optical axis.

This relation does not hold for pinhole cameras with just one hole. In this case, object and image distances can be chosen independently of each other.

We will investigate the relation between *g* and *b* more carefully.

We put an object at a large distance from the lens and seek the corresponding image distance. We then move the object even further away from the lens, and the image becomes smaller but stays focused. The image distance does not become smaller, Fig. 22.21. It has reached the smallest value possible. This value is called the *focal distance* (or focal length) *f* of the lens. The plane on which the image is produced is called the *focal plane*.

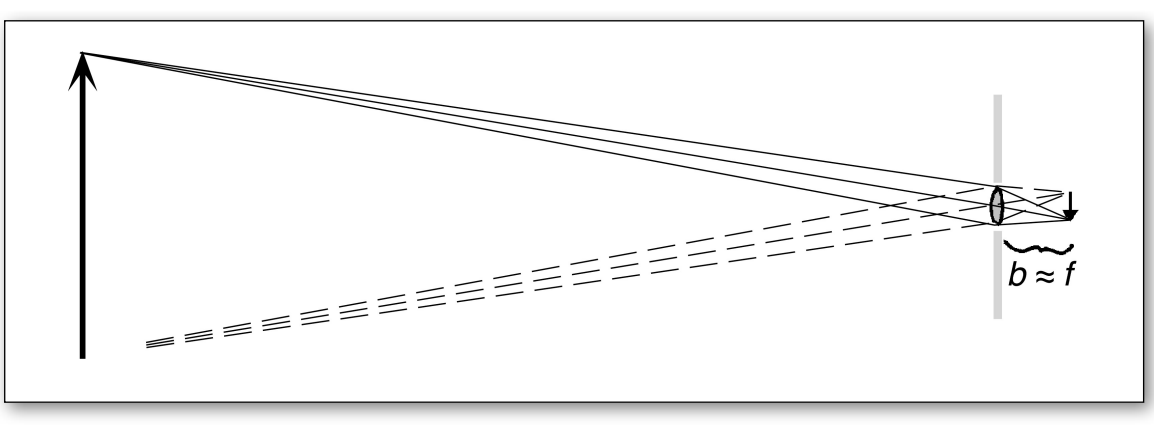

Fig. 22.21
The image distance is independent of the object distance if g is very large. It equals the focal length.

The following holds for objects that are very far away from a lens:

Image distance *b* = focal distance *f*

What happens if we shift the object in the other direction, closer to the lens? The image moves away from the lens. We cannot put the object too close to the lens because when the object distance approaches the focal distance, the image moves further and further away quickly. When *g* is smaller than *f*, there is no more focused image.

One can also say that:

When *g* = *f*, the image is at an “infinite distance”.

Or, symbolically:

If *g* = *f*, then *b* = ∞.

Fig. 22.22 shows two special cases.

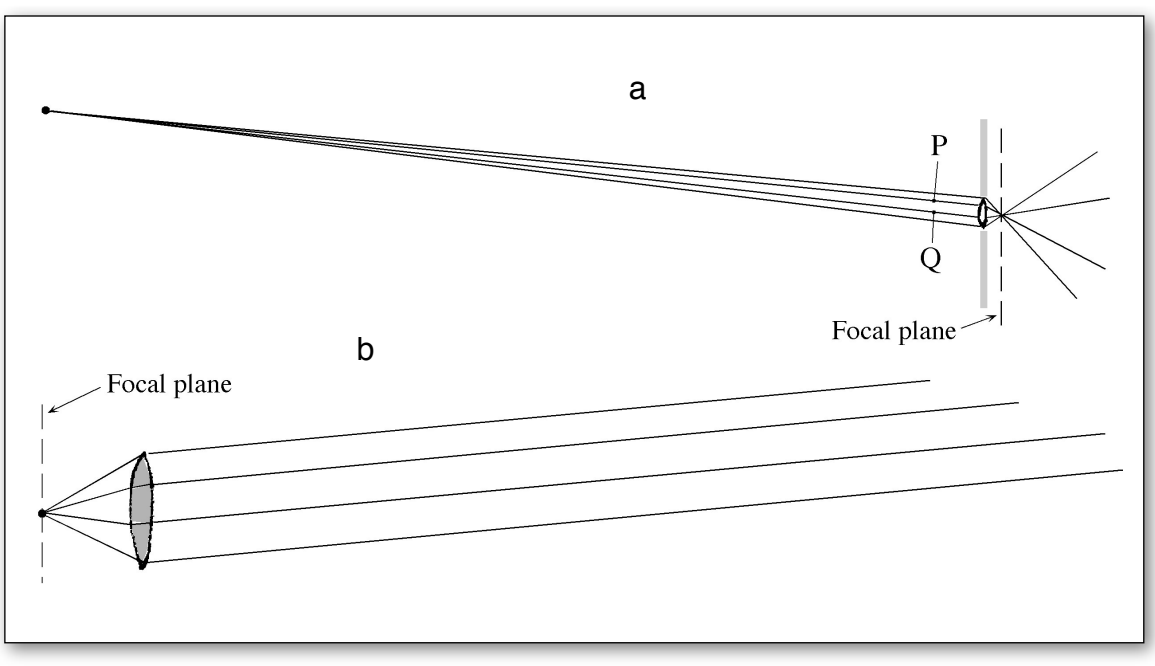

Fig. 22.22
(a) The object is a single luminous point at a large distance from the lens. The light behind the lens runs through a point on the focal plane on the right.
(b) The object is a single luminous point on the left hand focal plane. The light on the right hand side of the lens is parallel.

a) The object is a single luminous point at a large distance from the lens. The light falling upon the lens has just one direction at any point. The direction of the light at different locations, for example P and Q, is also the same. It is *parallel light*. The image of our object is a point on the focal plane on the right. In other words:

Parallel light falling upon a lens, continues behind the lens through a point on the focal plane.

b) The object is a luminous point lying on the focal plane. The image is a point at an “infinite distance”. The light on the right must, therefore, be parallel.

The light coming from a point-shaped source on the focal plane in front of a lens becomes parallel behind the lens.

Just as with a parabolic mirror, parallel light can be concentrated to a point and light from a point-shaped source can be made parallel.

Exercise

You have a lens with an unknown focal length. You also have a candle and some matches. How can you determine the focal distance of the lens? Describe two methods.

22.7 Focal distance and refractive power

Every lens has a characteristic focal length. Lenses can have very different focal lengths. How can we determine the focal distance of a lens? An optical image can be made where the object is at a great distance from the lens such as in Fig. 22.21. We measure the distance of the image from the lens, and because $f = b$, we obtain the focal distance. We use this method to determine the focal length of various lenses and find that:

> The stronger the outward curvature of a lens surface, the smaller the focal length.

The focal length is influenced by the curvatures of both surfaces. So the three lenses in Fig. 22.23 all have the same focal distance. As we see, a surface can even be curved inwardly when the other surface is that much more outwardly curved.

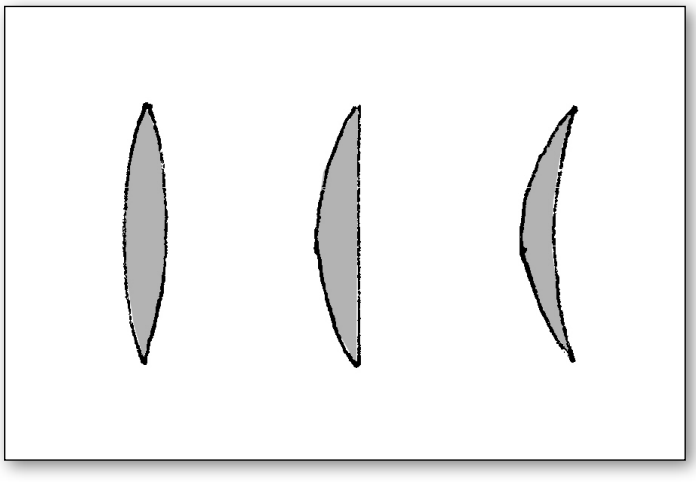

Fig. 22.23
All three lenses have the same focal distance.

Instead of describing a lens by its focal distance, the reciprocal is often used, the *refractive power D:*

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

The greater the refractive power of a lens, the more strongly the light rays at the edges of the lens are bent toward the middle.

The unit of refractive power is 1/m. This unit is also called a diopter, abbreviated to dpt. Therefore:

1 dpt = 1/m

A plane plate of glass that has (almost) no effect upon light has a refractive power of 0 dpt.

22.8 Combining lenses

We set up a lens L_1 with a focal length f_1 so that the image of a far away object is produced on a screen, Fig. 22.24a. Because the object is so far away, the image distance is equal to the focal distance of the lens.

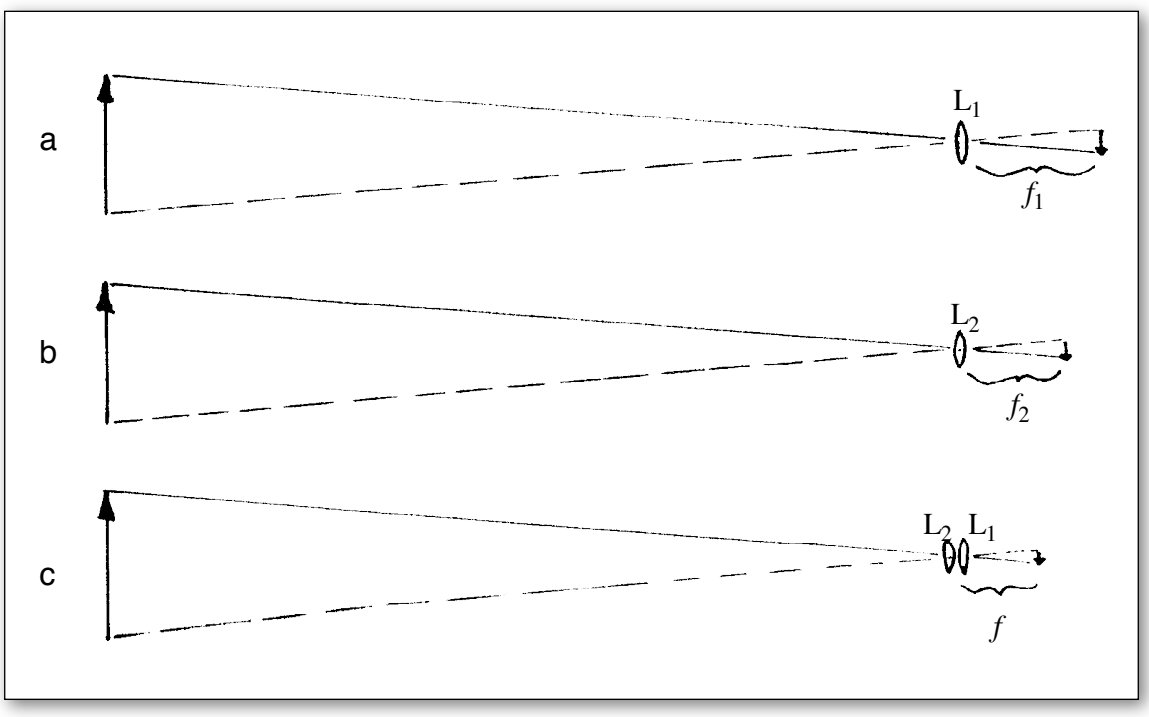

Fig. 22.24
The focal length of lens system (c) is smaller than that of lens L_1 (a) or that of lens L_2 (b).

We replace the lens by another one L_2 with a focal distance of f_2. The new image is at a distance f_2 from the lens, Fig. 22.24b.

Now we use both of them and put them as close together as possible. We again seek the place where the sharpest image is produced. We find that the image distance is smaller than either f_1 or f_2, Fig. 22.24c. (We take the image distance from the middle of the pair of lenses.)

The new image distance is the focal distance f of the lens system created by L_1 and L_2. Therefore

$f < f_1$ and $f < f_2$.

If we express it using refractive power, we say that the refractive power of the lens system D is greater than the refractive power D_1 of lens L_1 and greater than the refractive power D_2 of lens L_2. When the three focal distances f, f_1 and f_2 are measured, and converted to refractive power, we find that:

$D = D_1 + D_2$.

The refractive power of a lens system is equal to the sum of the refractive powers of the individual lenses.

On the left in Fig. 22.25 we see a lens L_1 which is not really a lens at all. It is thicker at its edges than in the middle. One of its surfaces is curved inwardly and the other is flat. We put this lens in front of another "real" one, lens L_2. The convex (outwardly curved) surface of the second lens fits perfectly into the concave (inwardly curved) one of the first lens.

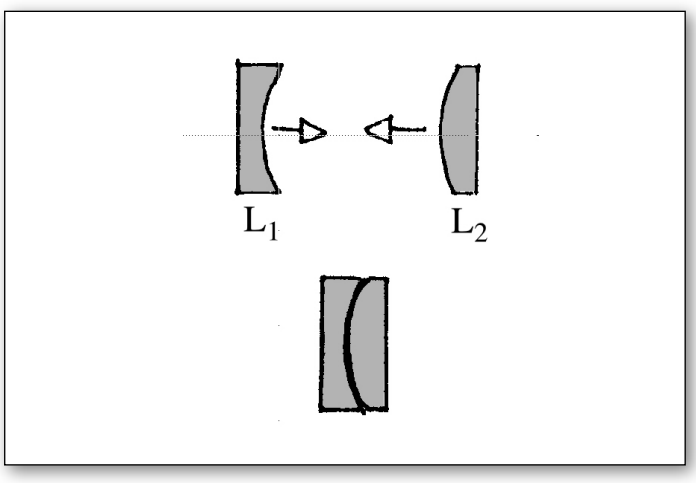

Fig. 22.25
The lens on the left has a negative refractive power, and the one on the right has a positive refractive power.

A coplanar plate (plane sheet) is produced, meaning a lens system with a refractive power of 0 dpt. We want the equation $D = D_1 + D_2$ to remain valid here as well. In order for the sum to correspond to a refractive power of 0 dpt, D_1 must be negative. If the lens on the right has four diopters, the one on the left must have minus four diopters:

$D_2 = 4$ dpt $\quad D_1 = -4$ dpt.

Lenses that are thicker at their edges than in the middle have negative refractive power.

If a lens with

$D_1 = 5$ dpt

is set before another one with

$D_2 = -1.5$ dpt,

the resulting lens system has

$D = 3.5$ dpt.

Exercises

1. A lens with a refractive power of 3 dpt and one of –4 dpt are combined into a lens system. What is the lens system's refractive power?

2. Two lenses, each with a focal distance of 40 cm, are put one behind the other. What is the refractive power of this lens system? What is its focal distance?

3. Three lenses are put one behind the other. The first one has a focal distance of 20 cm, and the second one has a focal distance of 50 cm. The third lens has a refractive power of minus two diopters. What is the refractive power of the lens system?

4. You wish to find the refractive power of a concave lens. You know that you need to do this differently than you would with a normal lens. What do you do? What equipment would you use?

22.9 Depth of field

Up to now, we have only dealt with optical images of flat objects, objects that can be described by one object plane. We run into problems, though, when an object is extended in the direction of the optical axis, meaning when it has an extension in depth.

We will consider a simple example of the optical image of two luminous points A and B that are at different distances g_A and g_B from the lens. Because g_A and g_B are not the same, the image distances b_A and b_B are not the same.

If b_A is the distance between the screen and the lens, Fig. 22.26a, then the picture of A is sharp. The light coming from B does not produce a sharp point but a larger spot. The picture of B is blurred.

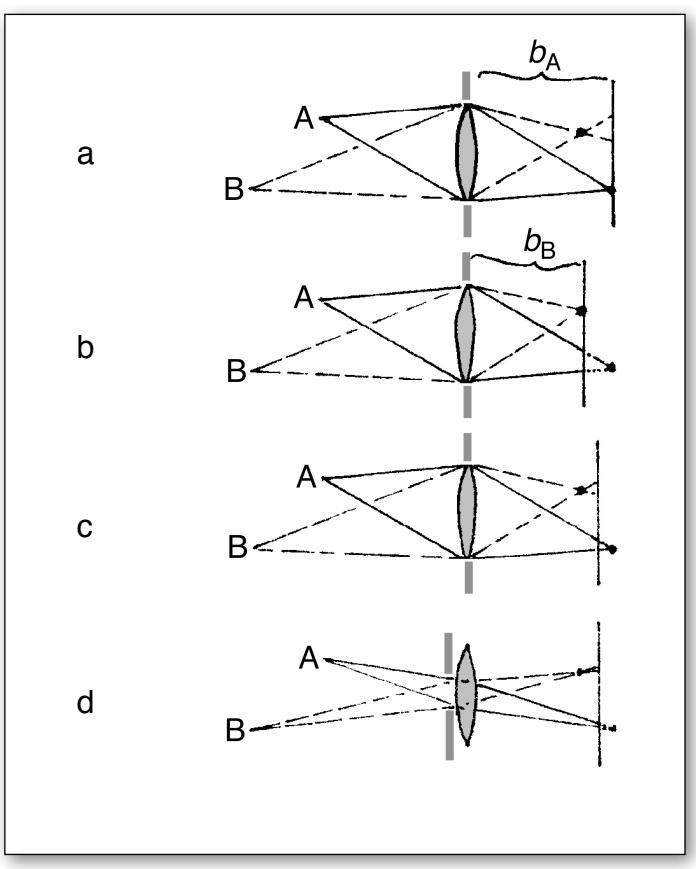

Fig. 22.26
(a) The image of A is sharp, the image of B is blurred.
(b) The image of B is sharp, that of A is blurred.
(c) A compromise: Both images are a little blurred.
(d) The diameter of the lens is reduced and both images become well focused.

If the distance between the screen and the lens is set at b_B, Fig. 22.26b, then the image of B is focused and that of A is blurred.

We see that when one picture is focused, the other is not.

We decide to compromise and put the screen in the middle between the two distances b_A und b_B, Fig. 22.26c. Now both images are blurred, but the image of A is not as blurred as in Fig. 22.26b, and the image of B is not as blurred as in Fig. 22.26a.

We will make one final alteration to our optical image: We put an aperture plate in front of the lens so that light can only pass through the very middle of it, Fig. 22.26d. It is as if we had replaced the lens with one of a smaller diameter (but the same focal distance).

What happens now? The picture becomes darker, but also sharper. We can see why this is by considering Fig. 22.26d. The spots (images) of A and B have become smaller. If the aperture is made smaller and smaller, the quality of the image will eventually be satisfactory.

There is always a certain range of depth of the side where the object is that is in focus. The smaller the diameter of the lens, the larger this area is. This range is called the depth of field.

We conclude that:

> The smaller the diameter of the lens, the greater the depth of field.

A smaller lens means less light and a darker image. Therefore:

> Large depth of field and great brightness of the image are mutually exclusive.

22.10 Objective lenses

You have now learned the basics of producing images with lenses. This is, by far, not everything there is to know. Maybe you have already noticed that a camera, a video camera, a video projector or over-head projector all use not one, but several lenses. They make use of an *objective lens*, or simply, an *objective.* (Notice that what we call an “objective lens” is actually composed of several lenses.)

As we have already seen, 2, 3 or more lenses can equal one lens. Shouldn’t it be possible to replace an objective with just one lens? We even know how to calculate the refractive power of this equivalent lens when we know the refractive power of the individual lenses making up the objective lens.

In spite of this, it is not the same when one lens is used instead of an objective. This is because what we have learned so far about optical images produced by one lens is only approximately correct. By closer examination, we see that a lens can never produce an exact image of a luminous point, but only an approximate image. Moreover, the image of a flat surface perpendicular to the optical axis is not really flat, but somewhat curved. Actually, a straight line results in an image that is slightly curved. Because of this, a white point is pictured as small colored rings. There are still other imaging errors or *aberrations*.

It is possible to reduce these imperfections by using several lenses in place of only one. However, calculating the effect of such an objective is a complicated matter.

Most modern cameras have objective lenses made up of at least 4 individual lenses. Fig. 22.27 shows a cross section of the objective of a camera. The objective lenses in microscopes are often composed of more than 10 lenses. Various types of glass must be used in producing these individual lenses.

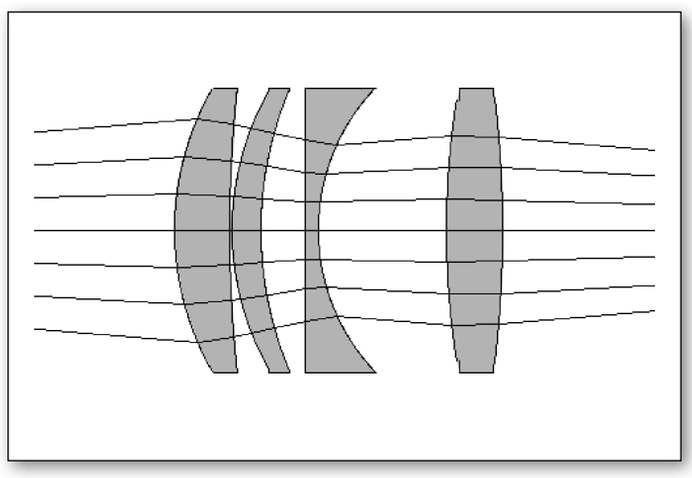

Fig. 22.27
A camera’s “objective lens” with four lenses

22.11 Cameras

We already have a basic idea of what a camera is. Fig. 22.28 shows the construction of one. The objective lens makes an image on the sensor of whatever is in front of the camera.

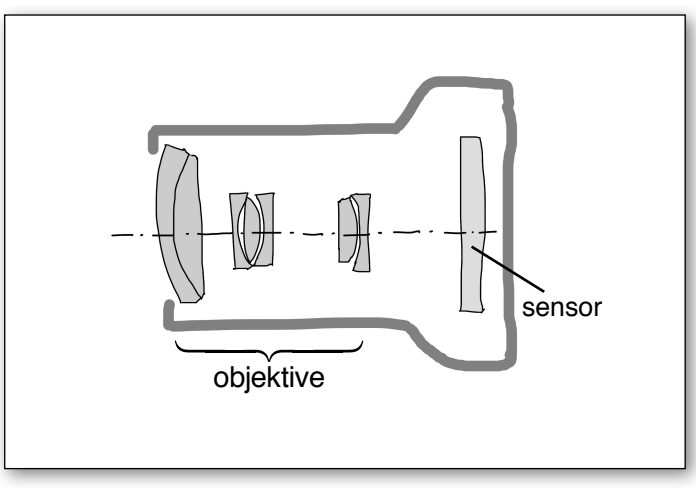

Fig. 22.28
Camera

In order for the "object" being photographed to be focused, the image distance needs to have the right value. This adjustment is usually done with a ring around the objective lens. When the ring is turned, the objective lens is moved relative to the film. The scale on the ring shows the distance of the object being photographed. This is why we say that we use the ring to adjust the distance of the object.

Normally, the light's path is obstructed by the *shutter*. In order to actually take the picture, the shutter release is pressed, opening the shutter for a very short moment. The length of time the shutter opens can be adjusted in most cameras. The shorter the amount of time the shutter is open, the less the chance of a wobbly picture. For objects that don't move too much, 1/60 second is sufficient.

Not only the distance and length of time for opening the shutter must be dealt with. An objective lens has a circular aperture whose diameter can be adjusted. The greater the size of the opening, the more light is allowed through the objective, and the smaller the depth of field is.

The numbers on the aperture scale show not only the aperture diameter but also the ratio

$$\frac{\text{focal length}}{\text{aperture diameter}} = \frac{f}{d}$$

Example: The largest aperture diameter of an objective lens with a focal length of 50 mm is 18 mm. The corresponding number on the aperture scale is
$50/18 \approx 2{,}8$.

This means that the higher the number, the smaller the aperture opening.

The three adjustments we just discussed—distance, exposure time, and aperture—are done automatically by many cameras. To do this, a camera needs to have a light meter as well as a range finder.

When the objects being photographed are always far away from the objective lens, the image distance is pretty much the same as the focal distance:

$b \approx f$.

Fig. 22.29 shows how the same object is photographed by three different objective lenses. The objective lens at the top of the Figure has the greatest focal length, the one at the bottom has the smallest. We see that the image of the object is larger, the greater the focal length (the smaller the refraction power) of the objective lens.

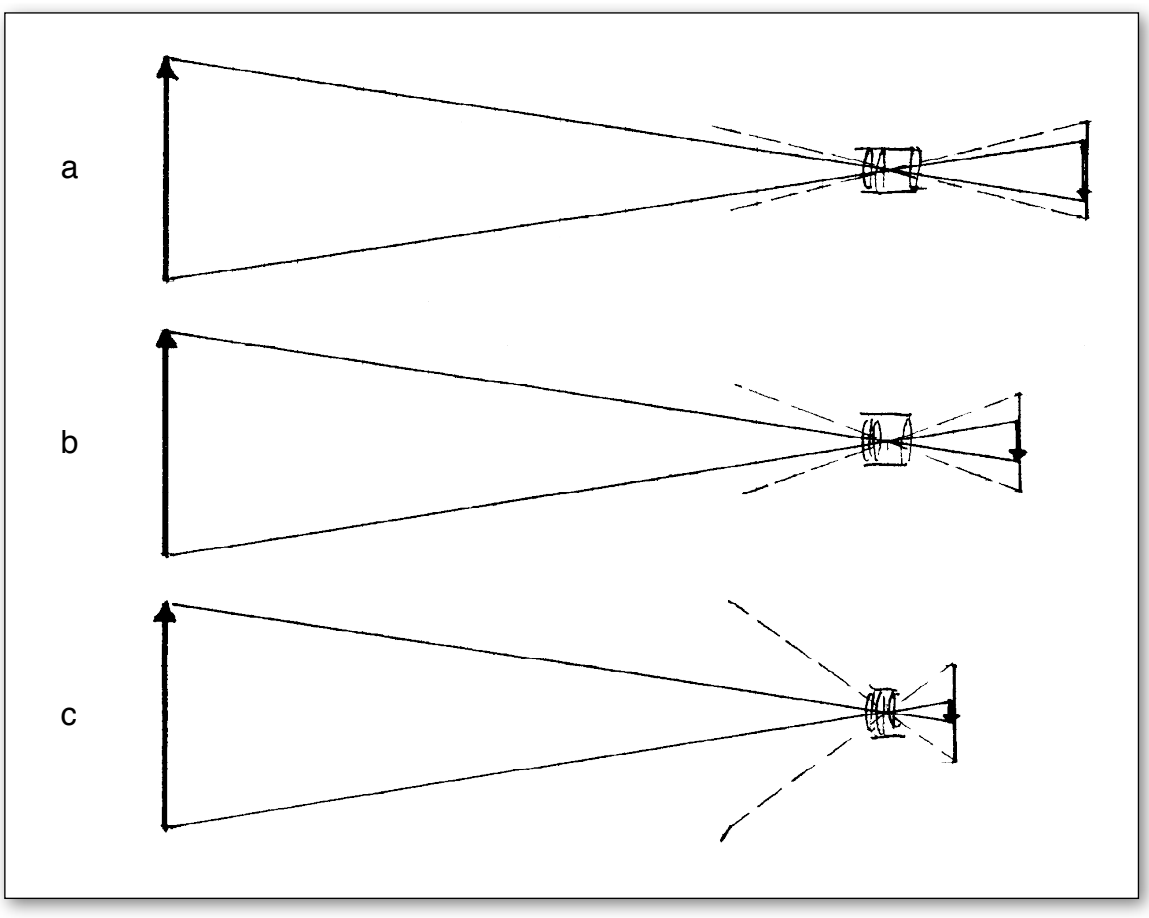

Fig. 22.29
The same object is photographed with lenses with three different focal lengths.
(a) a telephoto lens;
(b) a normal lens;
(c) a wide-angle lens.

When an object produces a large image on the sensor, this means that the angular range which the camera "sees" is small. It is possible to photograph a small angular range using an objective with great focal length and vice versa.

Objectives with great focal lengths "bring the object up close." These kinds of lenses are called *telephoto lenses*. Objective lenses with a small focal distance can photograph wide angular ranges. These are called *wide-angle lenses.* In between are the *normal objective lenses.*

We sum up:

> The greater the focal length, the larger the image of an object.

It is possible to exchange objective lenses (lenses with different focal lengths) on many cameras. An even more convenient method of changing focal distances is the zoom-lens. This is an objective with an adjustable focal length.

Exercises

1. When should the exposure time of a camera be kept short? When is a longer one necessary?

2. What is the advantage of a small aperture opening? What is the advantage of a large one?

3. When adjusting the distance you must take the aperture into account. Why?

4. A birthday party is taking place in a small room. You want to take a picture of all the party guests together. What kind of a lens do you use? Why?

5. You want to take a picture of your friend with mountains in the background. In the finder, you think the mountains look too unimpressive. You want them to make a big impression, so you will have to exchange the lens. What kind will you need? Why?

6. A 10 m high building is photographed from 200 m away. It is photographed once with a normal objective lens (f = 50 mm) and once with a telephoto lens (f = 180 mm). What is the height of the building on the sensor?

22.12 The eye

The eye is like a camera, Fig. 22.30. Light passes into the eye through the pupil. An image is produced on the retina. There is a lens behind the iris (the aperture that makes the pupil opening larger or smaller). The substance between the cornea and the lens is essentially water. The optic density of the material the lens is made of is somewhat higher than that of water. Refraction takes place mainly at the cornea, where light enters the eye (the refractive power is about 43 dpt). The lens has a smaller role in producing the optical image. Its most important function is “focusing”. By tensing, the orbicular muscle can increase the refractive power of the lens (from 15 to 27 dpt).

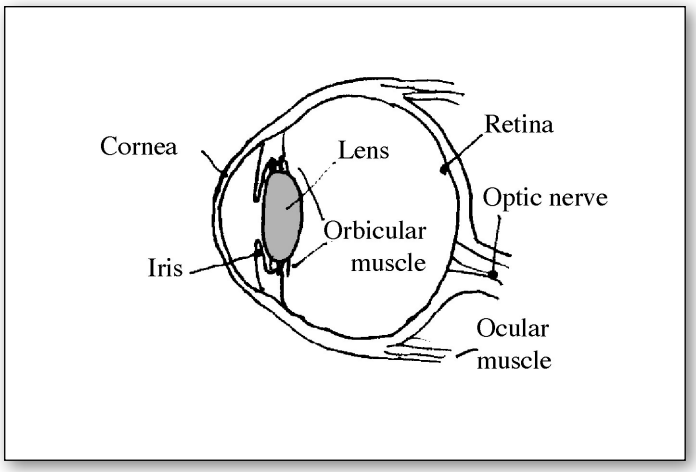

Fig. 22.30
The human eye

When the orbicular muscle is relaxed, the image of far away objects are sharp on the retina. Objects that are closer up become sharply focused when the muscle tenses up. The smallest distance that objects can be clearly seen at is about 10 cm. Try it out for yourself.

The retina is the actual data receiver. It is a receiver for data that comes into the eye with the carrier light. The retina contains a large number of individual receivers. The data passes from these, along the optic nerve, and into the brain.

22.13 Glasses and magnifying glasses

Glasses

If the orbicular muscle is relaxed, the image of a far away object should be in focus upon the retina.

The eyeballs of short-sighted people are too long. This means that the image of a far away object is produced in front of the retina. In other words, it is unfocussed and blurred, Fig. 22.31a. In order to obtain a sharp image, the refractive power of the eye's lens must be reduced. This cannot be done by tensing the orbicular muscle—that would only increase the refractive power and make the image even more blurred. Glasses having lenses with negative refractive power are the solution here, Fig. 22.31b.

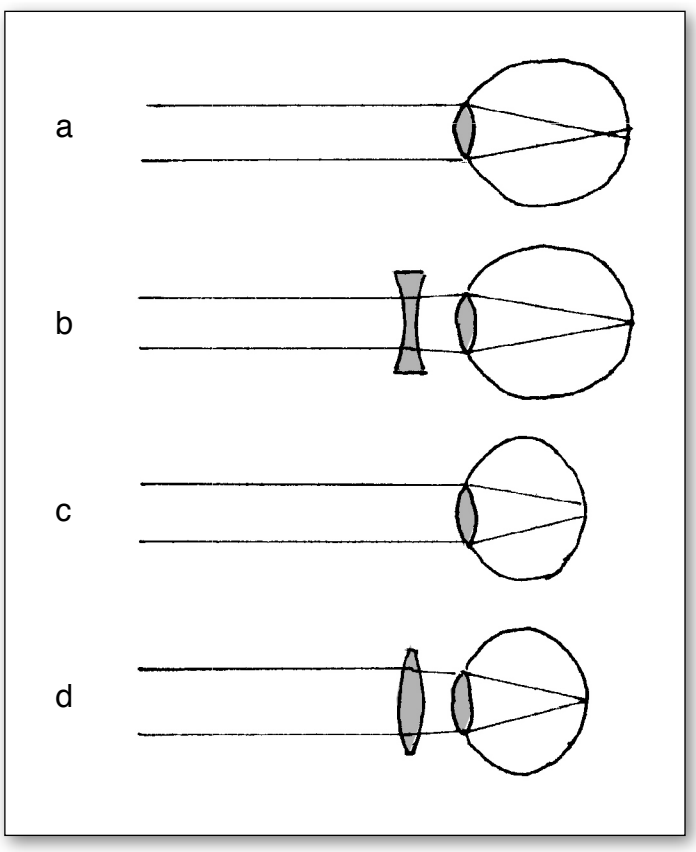

Fig. 22.31
(a) A short-sighted person's eyeballs are too long;
(b) correction using a lens with negative refractive power.
(c) A far-sighted person's eyeballs are too short;
(d) correction using a lens with positive refractive power

Far-sighted people have eyeballs that are too short, Fig. 22.31c. This results in the orbicular muscle needing to tense up in order for far away objects to be seen clearly. It also means that the smallest distance seen clearly is greater than that of people with normal vision. Far-sightedness can be corrected by lenses with positive refractive power, Fig. 22.31d.

As people age, the lenses in their eyes lose elasticity, and their refractive power cannot be changed as strongly by tensing of the orbicular muscle as it is with younger people. This means that close up objects cannot be seen as well as they once were. Glasses with positive refractive power can help here too. They are only necessary when the person is looking at an object up close, a book for example. When observing an object further away, the glasses must be removed, or … see Fig. 22.32.

Fig. 22.32
What is this man's vision problem?

Magnifying glasses

When we wish to see an object as clearly as possible, we often enlarge it by holding it as close to our eyes as we can—so close, that we just barely see it sharply. In order to enlarge it even more, it needs to be brought even closer. This would make the picture blurred because the refractive power of the eye is not great enough. We can help this situation by using a magnifying glass. A magnifying glass is nothing more than a lens with positive refractive power. The combined refractive power of the eye and the magnifying glass is greater than that of the eye alone. The magnifying glass has the same function as glasses for far-sighted people.

Exercises

1. How can you tell that glasses are for a shortsighted or farsighted person just by looking at them?

2. Try using your own or a borrowed pair of glasses to make an image of a burning light bulb upon a white wall. It could be that you don't manage to do it. Why not?

3. Magnifying glasses can be used as burning glasses. The lens concentrates sunlight onto one spot. What form does this spot have? Why does it have this form?

22.14 Video projectors

Let us imagine that projectors don't exist yet and we want to invent one. The simplest idea would be to take a small paper picture, light it as well as possible and project it onto a white screen using an objective lens, Fig. 22.33a.

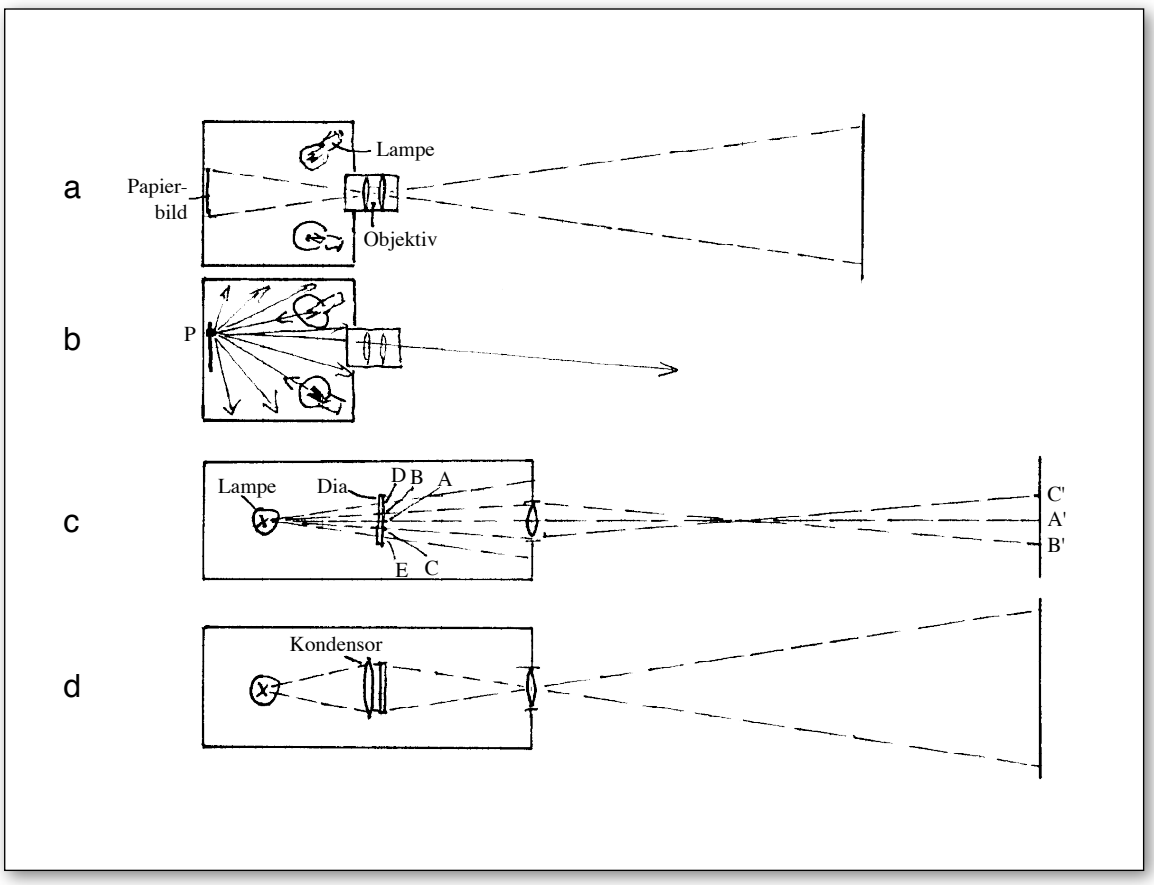

Fig. 22.33
(a) An episcope.
(b) An episcope loses a lot of light.
(c) The light passing through a slide is not scattered. Only the center of the slide is shown on the screen.
(d) The condensor bends all the light passing through it toward the objective lens.

These types of devices for projecting paper pictures actually do exist. They are called *episcopes*. You will see why they are not used very much. Fig. 22.33b shows where the light from the lamps falling upon point P of the image actually goes. Only a very small portion of this light passes through the objective lens and onto the screen. Most of it is absorbed at the inside walls of the projector. The reason for this is that paper scatters light. It is thrown back in all directions instead of going to the objective lens, where it should go.

Now we come up with another idea. We replace the paper picture with a picture that doesn't scatter light: We invent a slide. A slide either absorbs light or allows it to pass through, but does not scatter it.

Fig. 22.33c shows our first attempt to build a slide projector. At first glance, the projector looks pretty good. All of the light coming from point A on the slide passes through the objective lens and onto the screen. The same happens with the light from points B and C. However, things go wrong now. The light coming from points outside of B and C (D and E, for example) does not fall upon the objective lens, but misses it. What we see on the screen is a nice bright picture – but only of a circular part at the middle of the slide.

We now come up with a *condensor* which is a large lens. We put it close up to the slide, Fig. 22.33d. It bends the light rays coming from the lamp so that they all pass through the objective lens. As you can see in Fig. 22.33d, this means that the light source is projected onto the objective lens.

One last step is needed on the way to a modern projector: A bent mirror is put behind the lamp so that the light shining out the back of the lamp also passes through the slide to the objective lens. Finally, the slide is replaced an LCD matrix.

Exercises

1. A video projector should throw as much light as possible upon a screen. One might imagine that the objective lens would have the largest possible diameter, but this is not needed. Why not?

2. Episcopes always have objective lenses with large diameters. Why is this?

3. There is a screen 5 m away from a video projector. The projector should produce a 2.40 m high image. What must the focal length of the objective lens be? (The height of the LCD matrix is 24 mm.)

22.15 Microscopes

In order to make a large image of a small object, it must be brought up close to the focal plane of the lens making the image, Fig. 22.34a. The “device” in Fig. 22.34a is almost a microscope. In fact, the image produced by a microscope is not seen on a screen but directly with the eye. One looks down from above into the microscope.

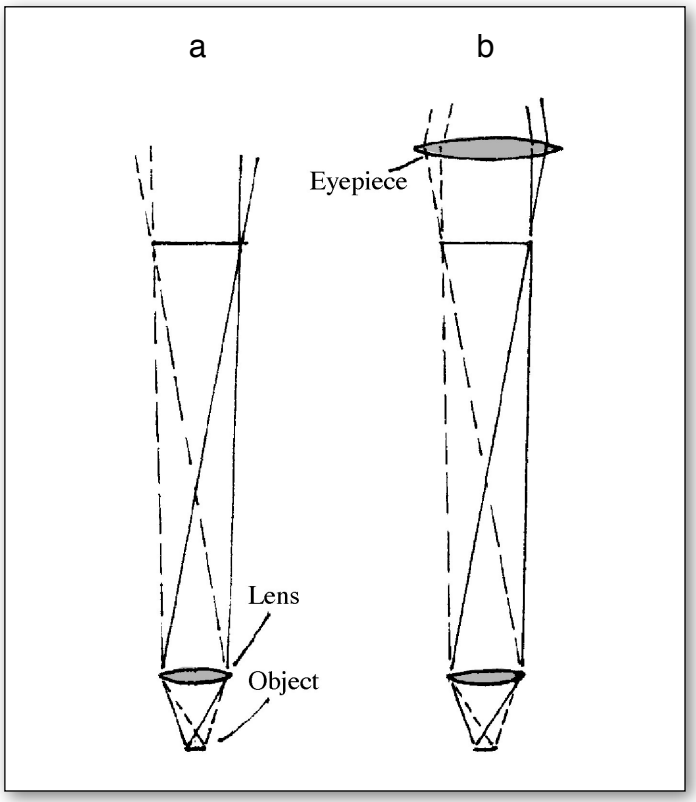

Fig. 22.34
A microscope.
(a) The object is close to the focal plane.
(b) The image is observed through a magnifying glass (the eyepiece).

The image should be as large as possible, meaning the eye should be as close as possible to the object. A magnifying glass is used to accomplish this, Fig. 22.34b. This kind of magnifying glass is an inherent part of every microscope and is called the eyepiece.

22.16 Telescopes

A telescope serves to make a large image of a far away object.

We have already seen that an image is the larger the greater the focal length of the lens or objective lens is.

A telescopic image is not produced upon a screen or film, but is directly observed from behind. In order to make the image as large as possible, it is seen through a magnifying glass or eyepiece just as with a microscope, 22.35.

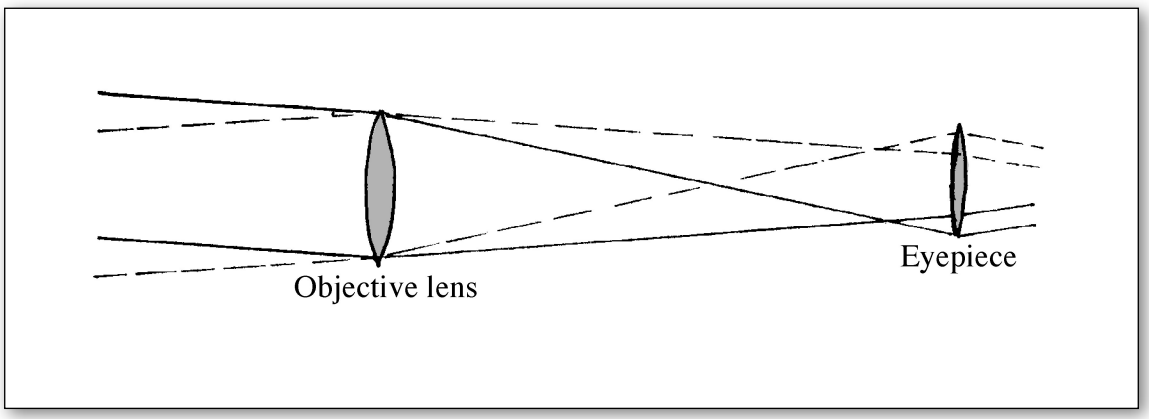

Fig. 22.35
A telescope. A lens with a large focal distance produces an image. This image is observed through a magnifying glass (eyepiece).

22.17 Astronomical telescopes

An astronomical telescope is basically nothing more than a large telescope or a camera's telescopic lens with a large focal length. It is built for observing the sky.

There is a special problem in observing stars. When the night sky is observed with the naked eye, one sees 3000 stars at most. This is only a tiny fraction of all the stars that exist. We don't know how many stars actually do exist. We cannot see most of them as they are too far away and too little of their light finds its way to us. Objective lenses with very large diameters are used to make those stars visible. The larger the diameter, the more light there is to produce the image.

It is possible to manufacture lenses with up to 1 m diameter only. For larger telescopes parabolic mirrors are used instead of lenses. A parabolic mirror produces an optical image in the same way a lens does.

The largest existing reflecting telescope has a diameter of about 8 m, and larger ones are being built.

The light coming to Earth from the universe is not only visible light. There are many kinds of invisible light as well. All of this radiation brings us interesting information about the composition of the universe and strange processes taking place in other stars and galaxies. For this reason it is interesting for us to study these different sorts of light.

There are special telescopes for each type of radiation (for each type of light). They make an optical image, and collect as much of the radiation in question as possible.

Radio telescopes are very similar to conventional optical telescopes. They collect radiation whose wavelengths are much larger than those of visible light. A simple grid of wires serves as the mirror for this radiation. For this reason, it is much easier to build very large parabolic mirrors for these kinds of radiation than for visible light. The largest movable radio telescope in existence has a mirror diameter of 100 m.

Exercises

1. The parabolic antenna for satellite television can be understood as a radio telescope. What can be observed by it and how?

2. The pupil of the human eye has a maximum diameter of 8 mm. How much more light does a large reflecting telescope with a diameter of 6 m collect than the eye?

23

Color

23.1 Three dimensional color space

In the following, we will discuss the color perception that light provokes in our eyes. We will not talk about a whole image on the retina but only one point of that image.

Color perception depends, of course, upon the kind of light in question. Generally, not just one pure type of light falls upon our eyes (upon one point on the retina) but a mixture of types of light. You might think that different light mixtures always evoke different kinds of color perception. This is not the case, though. There are a lot fewer color perceptions than there are light mixtures. In other words, many very different mixtures of light evoke the same perception of color.

For the moment we will only deal with how we perceive color and not the kinds of light causing it.

We will start by addressing a common problem: A certain color impression needs to be “transferred”. An example of this would be you trying to describe the color of your new tee shirt to a friend over the telephone. Or another example: The television sender trying to “communicate” to the receiver what the colors of the pixels on the screen should be.

We want to find out how such a message or transfer of data can happen. For clarity, we will first consider another situation where the same difficulty occurs.

You live in a small town and you want to explain to a friend in Japan where you live. You have, among others, the following possibilities:

1. You give the name of your town;
2. You give the longitude and latitude of your town.

If you use the first method, your friend will need to look for the name of your town in the register of an atlas. The coordinates there will allow him to find the town on a map.

The second method is simpler in that it is enough to give two numbers. There are two coordinate axes called longitude and latitude. If we say where the town is found on the two axes, we have described the location perfectly.

Now, back to color. You want to tell your friend the color of your new tee shirt. This problem is very similar to the one describing the location of your town. The list of color names is large: Yellow, red, blue, orange, ochre, pink, olive, and many more. Under every general color name there are more names that are less often used but that describe the color more precisely. Cadmium yellow, chromium oxide green, or sepia are some examples.

Instead of giving the name of the town, it is possible to give its coordinates. Instead of giving the name of a color, Fig. 23.1, it can also be described by giving coordinates.

Fig. 23.1
It is complicated to describe a color in words.

A small aside:

In order to tell someone how warm the day is, it is enough to say one thing: The temperature. Only a single “coordinate” is necessary.

To let someone know where we live, we need two pieces of information: Longitude and latitude. We need two coordinates in this case.

An airline pilot giving the tower information about the position of his airplane must give three pieces of information: Longitude, latitude and altitude. He uses three coordinates for this.

Back to color now. How many pieces of information must be given to describe a color? How many coordinates are needed to exactly specify a color perception?

You have probably owned a set of colored felt tipped pens and have tried to make order out of them. It probably did not work out very well. Violet – blue – turquoise – green – etc., many are easy to put in sequence but some are hard to find a place for. A dark blue one, for example, that doesn’t fit between blue and violet or between blue and turquoise.

This just shows that it is not possible to use one coordinate for all colors. In other words, color space is not single-dimensional.

You can try to use two coordinates to describe color, but it won’t work. You need three coordinates.

> Color space is three dimensional.

Color perception can be divided into three different aspects. These are *hue, saturation,* and *brightness.*

Imagine a large number of differently colored cubes in front of you. Every clearly distinguishable color is there.

First we take out all the cubes with strong, intensive colors. We put these into a series that, interestingly, can be put into a circle, Fig. 23.2.

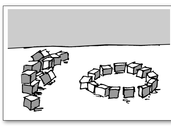

Fig. 23.2
All the strong, intensive colors can be arranged into a closed scale or color circle.

This circle is called a *color circle.* The cubes differ in their hue. Of course, the colors continuously merge into each other in the circle. We give twelve of the colors names, Fig. 23.3:

Red – orange-red – orange – yellow – yellow-green – green – turquoise – cyan – blue – blue-violet – purple – magenta

Fig. 23.3
The color circle

We could just as well have used the numbers 1 through 12 on this color circle.

We look through our pile of cubes and find still more of them that go with these hues. For example, with the intensive blue cube we can find others in the pile that are paler *or* darker. We also find some that are paler *and* darker. “Pale” and “dark” are not mutually exclusive concepts here.

Only the bright ones among these blue cubes can again be put into a series. At one end is the strong blue cube, in between are the pastel blue ones, and at the other end is a white one. We see that starting from every strong color, there is a continuous transition towards white. One says that the saturation changes in the process of going from strong color to white. Strong colors are saturated and white is totally unsaturated. We put the pale colored cubes inside the circle of strong color toned ones so that the saturation decreases towards the middle. In the exact center is the white cube. We now have a color disk, Fig. 23.4.

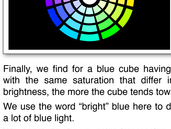

Fig. 23.4
The color disk. The colors on the edge are intensive and become paler toward the middle.

Finally, we find for a blue cube having a certain saturation others with the same saturation that differ in brightness. The less the brightness, the more the cube tends toward black.

We use the word “bright” blue here to describe a surface that emits a lot of blue light.

We conclude that a color impression can be described by three characteristics. These are “hue”, “saturation” and “brightness”. This means that we can create a three dimensional object out of our color cubes. It looks a bit like a cylinder, Fig. 23.5. The bright, saturated colors are on the upper, outer edge. The saturation decreases towards the middle, and right at the top in the center is white. The brightness decreases downward. At the top of the cylinder axis is white that gradually becomes a darker and darker gray towards the bottom. The bottom most surface of the cylinder is black. This cylinder represents a three dimensional model of *color space.*

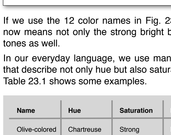

Fig. 23.5
The color cylinder is a model of three-dimensional color space.

If we use the 12 color names in Fig. 23.3 to describe hues, “blue” now means not only the strong bright blue, but pale and dark blue tones as well.

In our everyday language, we use many different names for colors that describe not only hue but also saturation or brightness (or both). Table 23.1 shows some examples.

Name	Hue	Saturation	Brightness
Olive-colored	Chartreuse	Strong	Middle
Beige	Yellow	Light	Middle
Pink	Red	Light	Strong
Brown	Orange	Strong	Light

Table 23.1
Hue, saturation and brightness of some well-known colors

You should now be able to describe any color by its hue, saturation and brightness—even “murky” or “in between” colors.

> Color impressions have three different characteristics: Hue, saturation and brightness.

Exercises

1. Color space can be represented in other ways than by a cylinder. How might you do this?

2. Discuss the question of the beginning and end of a scale for hue, saturation and brightness.

3. The scale for hue is a closed scale. Name another quantity whose values lie upon a closed scale.

4. Describe the colors of the following objects by qualitative indication of hue, saturation, and brightness:

Bread, cocoa powder, chocolate milk, cola, an artichoke, your skin, a zinc roof gutter, rust, the floor, walls and ceiling of your classroom, train cars.

23.2 Mixing light

We experiment with two slide projectors. However, instead of putting in slides, we put in color filters. Each projector produces a colored square upon the wall. We orient the projectors so that the squares overlap partly, Fig. 23.6.

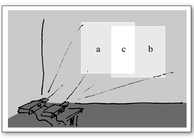

Fig. 23.6
The bright squares projected onto the wall overlap in area c.

The light from one projector is backscattered from area a, the same happens with the second projector's light at area b. From the area where they overlap (c), both types of light are backscattered simultaneously. We call the colors of the three types of light A, B, and C. The light coming from c is a mixture of the lights coming from a and b.

We will use the projectors to once again show what the terms brightness and saturation mean.

A green filter is inserted into each projector. Colors A and B are a saturated green. What color C does the light coming from area c have? The hue and saturation of C must be same as A and B. However, twice as much light is coming from each square centimeter of c as comes from a or b. This means that C differs from A or B by its brightness, Fig. 23.7.

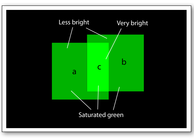

Fig. 23.7
The brightness in the area of overlapping (c) is greater than in a or b.

We now remove one of the filters. White light now gets from that projector to the wall. The light in area a is a saturated green, and the light in b is simply white. The overlapping area c is a pale green. The color C is a weakly saturated green. C is also brighter than A or B, Fig. 23.8. The saturation of a color can be reduced when white light is mixed into it.

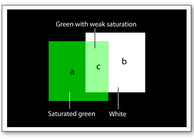

Fig. 23.8
The green in a is saturated. The green in the overlapping area c has weak saturation.

Now we will try a somewhat more complicated mixture. Different color filters are inserted into the projectors. We start with a red filter in one and a yellow one in the other. Color A is red, color B is yellow, and color C is orange, Fig. 23.9.

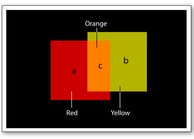

Fig. 23.9
Although two types of light come from the overlapping area, only one color impression is produced.

This is no surprise to you, but something is worth thinking about here. In area c, only one color appears, namely orange. We do not see the colors red and yellow individually. Our eyes "work" differently than our hearing does. When we hear music coming out of a loudspeaker, we can pick out the individual instruments playing it.

For the following experiment we will only use filters that produce a saturated color as long as their light is not mixed with other sorts of light.

It turns out that the color mixture resulting from two types of light is easy to predict using the color disk.

Two arbitrary color impressions A and B are connected to each other on the color wheel by circular arcs, Fig. 23.10. If A and B are not opposite each other, the arcs are of unequal length. The hue of the mixture of A and B is always a color tone of the shorter arc. The color mixture is saturated when A and B are very close to each other, or when the arc is very short.

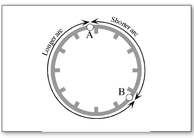

Fig. 23.10
The two colors A and B are connected on the color circle by two circular arcs.

The longer the shorter arc is, the less saturated the mixed color C is. If A and B lie across from each other, the mixture is totally unsaturated: It is white.

Examples: Orange and yellow-green lie close to each other. The mixed color is a rather saturated yellow, Fig. 23.11a. Orange-red and green are further away from each other and the result of mixing them is again yellow, but much less saturated than before, Fig. 23.11b. Red and turquoise are even further apart. Colors cannot be further apart than in this example. The color mixture is totally unsaturated, meaning it is white, Fig. 23.11c.

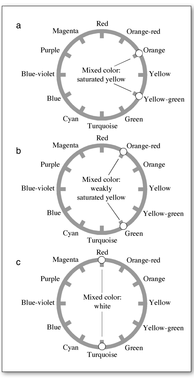

Fig. 23.11
(a) Orange + yellow-green = saturated yellow;
(b) orange-red + green = weakly saturated yellow;
(c) red + turquoise = totally unsaturated yellow = white

When color B is mixed with color A and the result is white, one says that B is the *complementary color* to A. Blue-violet is the complementary color to yellow, magenta is the complementary color to green, and red is the complementary color to turquoise.

The fact that colors lying across from each other on the circle result in white doesn't happen by chance. The exact position of colors on the circle has been chosen so that the complementary colors lie opposite each other.

Now we will try a still more complicated mixture of light. We will use three projectors so that we have three overlapping areas of two colors each, and one area where all three colors overlap, Fig. 23.12.

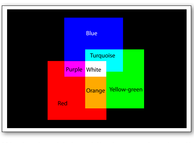

Fig. 23.12
Here we have three overlapping areas of two colors each, and one overlapping area of three colors.

We try to obtain white by overlapping three colors. Maybe you already know how to do this. The three individual colors must form the corners of a equilateral triangle on the color circle, Fig. 23.13. Here are two examples:

Red + yellow-green + blue = white

Purple + turquoise + orange = white

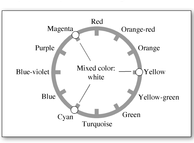

Fig. 23.13
Three colors that form an equilateral triangle on the color circle produce the mixed color white.

Correspondingly, it is possible to produce white with four projectors when the four colors form the corners of a square (a rectangle does it too). Even more projectors can be used. Whenever the individual colors on the circle form an equilateral polygon, the resulting color mixture is white. This is actually how the white light of the sun, a light bulb or a fluorescent tube is created. All of these light sources emit a mixture of many types of light whose colors lie at the edge of the color circle.

23.3 How the eye can be deceived – television images

We made mixtures of two kinds of light, for example the ones that, by themselves, make the color impressions blue and green. It happened like this: blue and green light are emitted at the same time, and from the exact same spot on the screen. The resulting color impression was turquoise.

It is possible to mix blue and green light less carefully, and still create the color impression of turquoise. The eye can be deceived. There are two ways to do this.

The components of the light do not arrive simultaneously.

A disc is divided into colored sectors. In a dark room, it starts to spin, and a powerful white beam of light shines upon it, Fig. 23.14. We do not see the individual colors any longer. The disk appears in a uniform color, the mixture of the colors on it. We find that if we have one red sector, and one yellow sector, we see orange. If we have three sectors, namely red, yellow-green and blue, we will see white (actually, a dark white or gray).

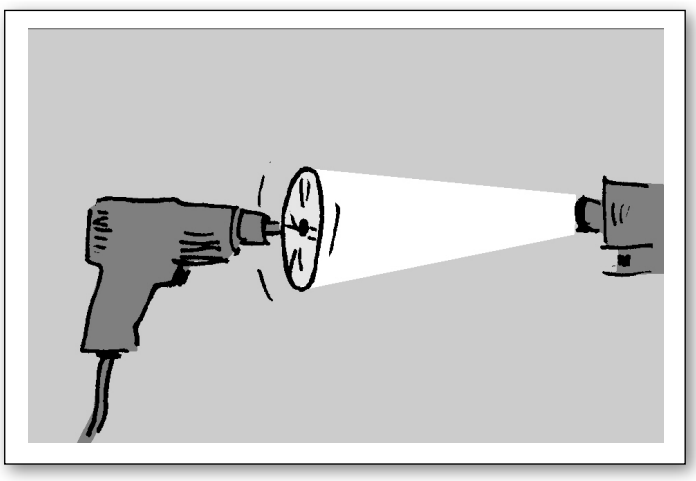

Fig. 23.14
When the disc spins very quickly, the colors of the different sectors produce one mixed color for our eyes.

The light components do not come from the same spot.

The green and blue light do not have to come from the same spot. This also means that the two kinds of light do not need to fall upon exactly the same spot on the retina. It is enough if the places the green light comes from lie so close to the spots the blue light comes from, that the eye cannot “resolve” them anymore.

This possibility is often exploited. Television and computer screens are made up of a grid of small spots called *pixels*. There are three sorts of pixels: they glow red, yellow-green and blue, Fig. 23.15. Take a close look at a television screen using a magnifying glass if possible.

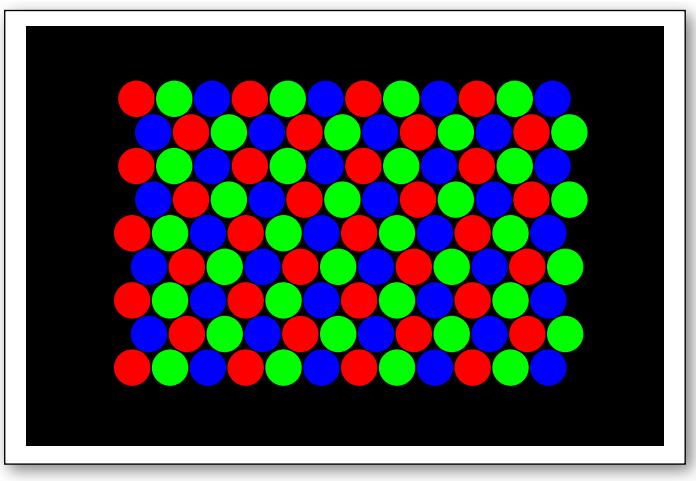

Fig. 23.15
A section of a television screen. There are three kinds of pixels.

From a distance it is not possible to see the individual colors when all the pixels are lit but only a uniform white. If just the red and yellow-green pixels are lit, one sees yellow, etc.

The three types of pixels can produce every color impression. They can not only be turned on and off. Each pixel can also glow more or less brightly. We will look at a few examples.

The red and blue pixels of a screen are dark. The screen is yellow-green. If the brightness of the red pixels is slowly increased, the screen will gradually become yellow. When the red pixels are at their brightest, the screen becomes somewhat orange. Now if the brightness of the yellow-green pixels is slowly reduced, the color of the screen changes toward red. When the yellow-green pixels are finally turned off, the screen is totally red.

In this way, red can move through purple to blue, and from blue through turquoise back to yellow-green.

Pale colors are obtained by illuminating not only two kinds of pixels, but by allowing the third to glow as well. If all three kinds of pixels are at maximum brightness, we see white.

Exercises

1. How does the brightness of the three types of pixels need to be adjusted so that the following colors are produced? Yellow, violet, pink, olive, ochre, and dark gray.

2. We stated that all colors can be produced on a television screen by using the three kinds of pixels. This statement is not quite accurate. Why not?

23.4 Back to color space

Let's take another detour here. We wish, again, to describe to someone the location of our town. You remember that it is enough to give the longitude and latitude, meaning just two numbers. Let us now assume that the entire country is covered by a system of streets that form a quadratic grid. The streets do not run north to south and east to west, but diagonally: from southwest to northeast and from southeast to northwest, Fig. 23.16. The streets make up a coordinate system of right angles. The origin is the country's capital city.

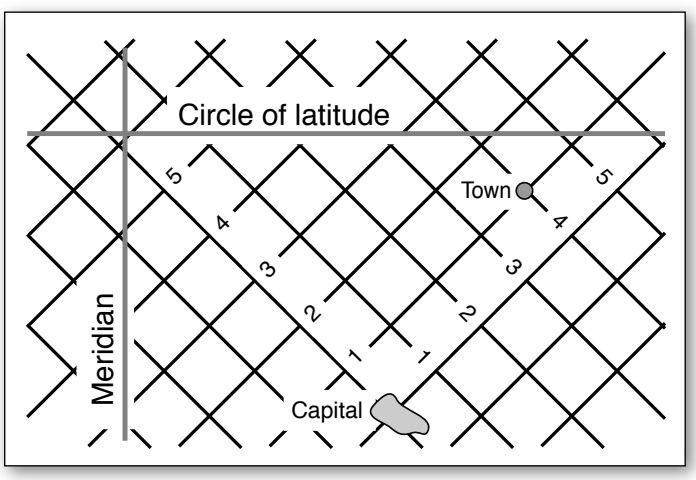

Fig. 23.16
The streets make up a system of coordinates that is at a 45° angle to meridians and circles of latitude.

We can now indicate the location of our town differently than before. As before, we need to give two numbers: The northeast coordinate and the northwest coordinate. We come to an important conclusion here. Independent of the kind of coordinate system, two numbers are always needed. Fig. 23.17 shows still another possibility. In this case, the distance r from the origin and the angle α to the axis g are given to describe the town's location. Again, it is two numbers.

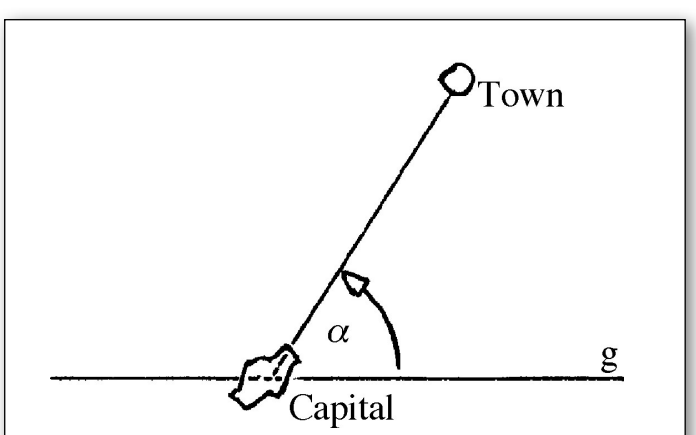

Fig. 23.17
In order to determine the location of a town, the distance to the capital city as well as the angle with respect to g are given.

We have a similar situation with color, but in this case, we need to give three coordinates. To be exact, we need three *numbers*. We have not worried about a precise calibration of the color coordinates, so we give only an approximation of the location of the color impression on the three coordinate axes.

Just as for geographic location, we can choose the coordinate system for color impressions differently. Moreover, it makes no difference what system is chosen—three values are always necessary to describe a color impression.

We have already gotten to know two examples of different color-coordinate systems:

1. Hue – saturation – brightness;
2. Brightness of red – brightness of yellow-green – brightness of blue.

In Case 2, three primary colors are used to describe a color impression. These are red, yellow-green, and blue.

You can probably imagine what some other color-coordinate systems might look like. Three other colors can be used as the primary colors that create a equilateral triangle on a color circle, for example:

Orange – turquoise – purple.

The brightness of these three colors is indicated on the coordinate axes.

Exercise

Assume that a television screen works with the pixel colors orange—turquoise—purple (instead of red—yellow-green—blue). How does the brightness of the pixels need to be adjusted to create the following color impressions? Red, blue, pink, white, brown, black.

23.5 Spectra

In this section we will not discuss the color impression that light produces in our eyes, but the composition of light itself.

Light is commonly composed of several kinds of light, both visible and invisible.

We are now interested in what light is made up of. It doesn't matter that we can see with the help of light, or that we can perceive color with it.

Our first problem: We wish to describe the composition of a certain mixture of light. Maybe you think that you can do this already. We know that three pieces of information are necessary for this. Wrong! We can describe the color impression in our eyes with three values, but not the mixture of light. Remember that many different light mixtures can produce the same color impression in our eyes.

In order to make clear how a light mixture can be characterized uniquely, we will first look at another, related, problem.

We wish to make a graphic representation of the age structure of the population of a city. Fig. 23.18a shows one possibility for doing this. The people are put into age groups: 0 to 20 years old, between 20 and 40 years old, etc. Everyone falls into one of these categories.

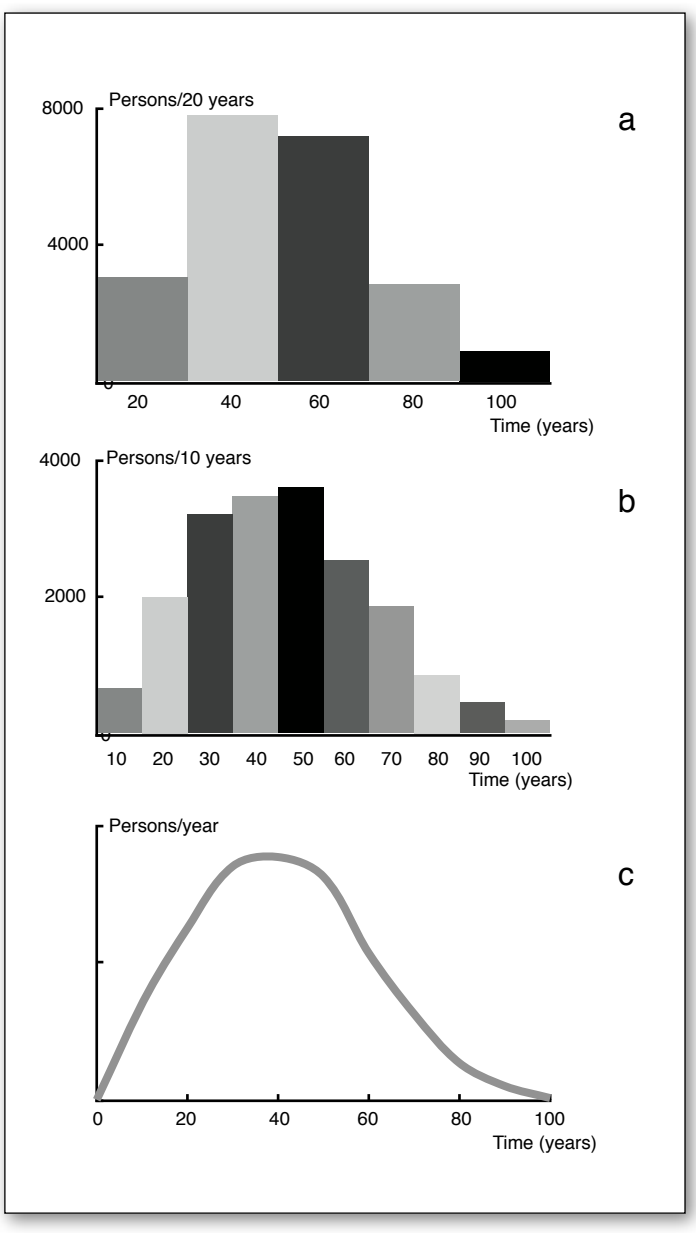

Fig. 23.18
Graphic representation of the age structure of a city.
(a) Twenty year intervals;
(b) Ten year intervals;
(c) Continuous representation.

You see right away that this representation does not show much detail. It is impossible to tell high school students from babies. This can be fixed by choosing a more refined classification, Fig. 23.18b. If this is still not enough, it can be refined even more. Finally, the steps become so small that it is possible to draw a smooth line through the population of every age group, Fig. 23.18c. This curve is called an age spectrum.

Our problem with light is very similar to this. We remind ourselves that light is always moving and that when we are dealing with light we are dealing with currents.

Light is made up of components with different wavelengths. (Remember that "visible" light has a wavelength of between 400 nm and 800 nm.)

In order to represent the composition of a certain light mixture, we divide the wavelengths into ranges, for example 20 nm. The bar in Fig. 23.19a for each range indicates how much light the mixture contains from that range. We use the energy current as the measure of the light current (the number of Joules that the light of that wavelength range transports per second).

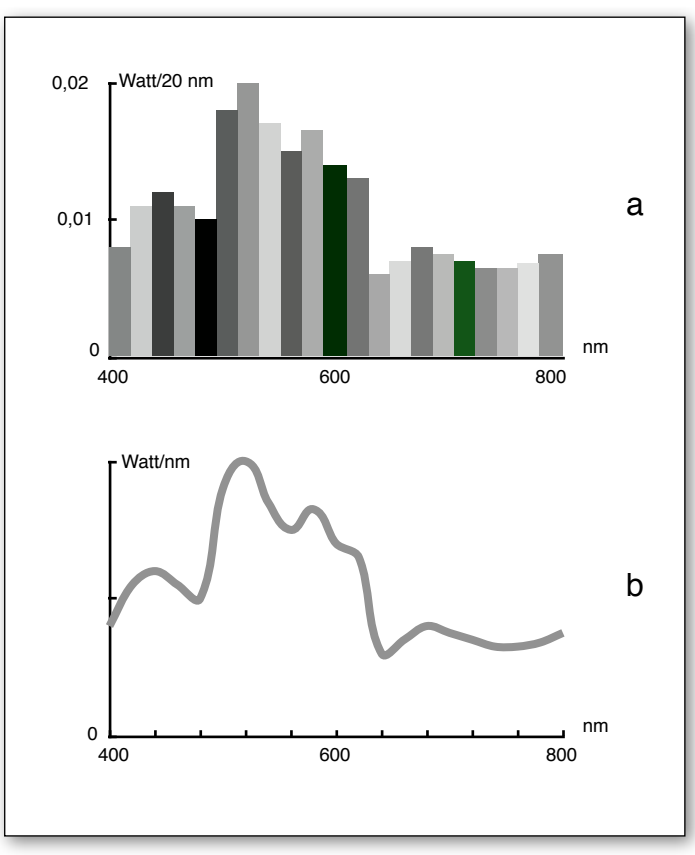

Fig. 23.19
Graphic representation of the composition of a light mixture.
(a) 20 nm intervals;
(b) Continuous representation.

It could be that the graph is too coarse. We refine it until we obtain a continuous curve which we call a spectrum, Fig. 23.19b.

Instruments used for recording spectra are called *spectrometers*. Prisms are used in many spectrometers for analyzing light.

A rough impression of the spectrum of a certain light mixture can be obtained by using a prism to separate the different light rays into different directions and shine them upon a screen. We will look at the spectra of different light sources. Fig. 23.20a shows the spectrum of sunlight, Fig. 23.20b shows that of a light bulb, and the spectrum in Fig. 23.20c is the spectrum of a sodium vapor lamp. The spectra of sunlight and light bulbs contain light of every wavelength. The sodium vapor lamp, on the other hand, mainly contains light of one wavelength: *monochromatic light*. There are also many other kinds of lamps that emit light of only one or just a few wavelengths. These are called spectral lamps.

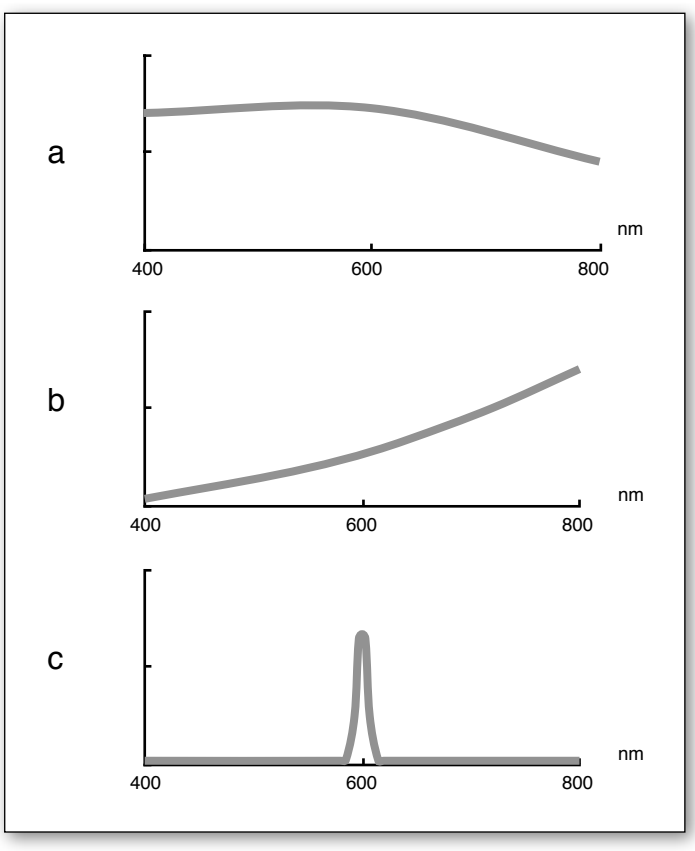

Fig. 23.20
Spectra. (a) Sunlight; (b) light bulb; (c) sodium vapor lamp

23.6 The relationship between spectrum and color impression

Again, we consider the color disk which is the top of the color cylinder. All the colors are very bright. The saturated colors lie at the outer edge. Toward the middle they gradually become less saturated. In the exact center we find white. What spectra belong to the various color impressions of the color disk?

Monochromatic light always produces color impressions with maximum saturation.

The reverse, namely that saturated color impressions are produced by monochromatic light, is not always true. It is valid for the colors marked with the dashed line in Fig. 23.21.

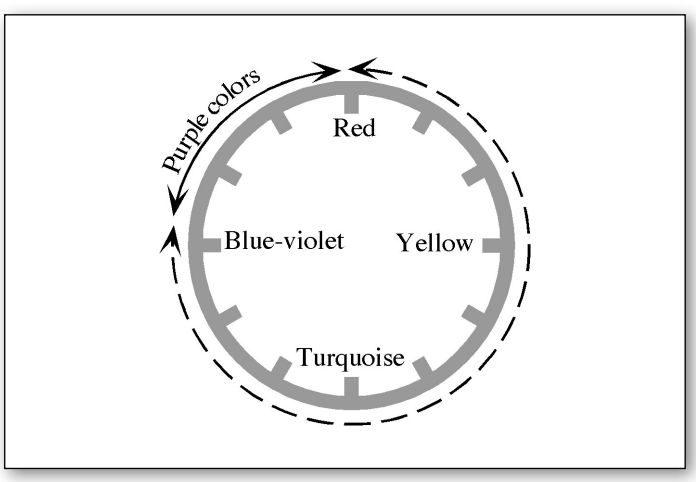

Fig. 23.21
Only the colors on the color circle that are marked with the dashed line belong to just a single type of light.

Every color that does not lie upon the edge of the color disk can be produced by more than one spectrum. The less saturated a color is, the more possibilities there are. You have already seen this in the case of white. For example, white can be produced by mixing two types of light that lie at opposite points on the color circle.

We see that with the exception of the colors lying at the edge of the color disk, many different spectra belong to every color.

We can only perceive a small fraction of the complexity of a spectrum with our eyes. This is an example of *data reduction*. A lot of data enters our eyes with a spectrum. Much less is actually relayed to the brain. Instead of the entire spectrum, we could say that only three numbers are transferred. (Don't forget that we are talking about only one so-called pixel here, meaning only one point on the retina.)

We have avoided one problem so far. The saturated colors between red and blue-violet – purple, magenta.... – do not correspond to any monochromatic light. These are called the purple colors. Above red on the wavelength scale the light is not purple but infrared, meaning it is invisible. Below violet, it is also not purple but ultraviolet, which is invisible as well.

In order to produce saturated purple colors, *two* types of pure light must be mixed: red light and violet light. Whether the color impression will be closer to red or violet depends upon the proportion of each type of light.

Exercises

1. Sketch two different spectra that belong to the color impression pale yellow.

2. Sketch three different spectra that belong to the color impression white.