KPK

The Karlsruhe Physics Course

for the secondary school A-level

Oscillations, Waves, Data

The Karlsruhe Physics Cours

A textbook for the secondary school A-level

- Electrodynamics
- Thermodynamics
- **Oscillations, Waves, Data**
- Mechanics
- Atomic Physics, Nuclear Physics, Particle Physics

Herrmann

The Karlsruhe Physics Course

Issue 2019

Edited by Prof. Dr. *Friedrich Herrmann* and *Dr. Holger Hauptmann*

Translation: *Kathrin Schilling*

Layout: *H. Schwarze*

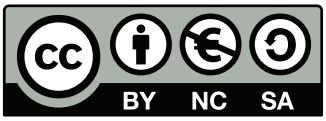

TABLE OF CONTENT

5 Interference of Light and X-Rays

6 Data Transfer and Storage

1 OSCILLATIONS

1.1 Provisional description

Oscillations are processes that are particularly important for the physical description of the world.

The best-known example of an oscillation process is the movement of an object that is suspended on a thread or on a rope, Fig. 1.1.

Another example can be realized even more easily. A ruler is clamped on one end while the other end is pushed slightly downwards and released, Fig. 1.2. The non-clamped end is shaking up and down. Besides that, there are many other examples.

To see the essential aspects of an oscillation in the sense of physics, we will at first examine a system that is slightly simpler than the pendulum and the ruler but that might appear somehow unnatural to you, Fig. 1.3.

Two bodies A and B are connected to each other by an elastic spring. They can only move back and forth in a single direction. We assume that the movement is not slowed down by friction. At first, the spring is not tensioned and the bodies are not moving yet. Then we displace A slightly to the left and B to the right by the same distance, and release them. Both bodies perform a back-and-forth movement, i.e. they „oscillate“ against one another. When A moves to the right, B moves to the left and vice versa.

Studying this arrangement, we would like to understand how the process becomes an oscillation.

Even superficial examination tells us: once initiated, the process continues by itself. „Initiated“ can be formulated more precisely. We have extended the spring and therefore charged the system with energy. As we have excluded friction and as there is no other outlet for the energy either, the energy remains trapped in the oscillating system.

Notably, the process is periodic. If we take a closer look at the oscillations, we will find that the duration of an entire period, the *oscillation period T*, is

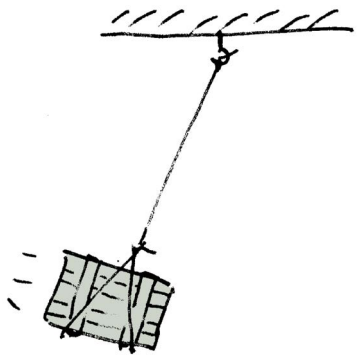

Fig. 1.1 Oscillation of a suspended body

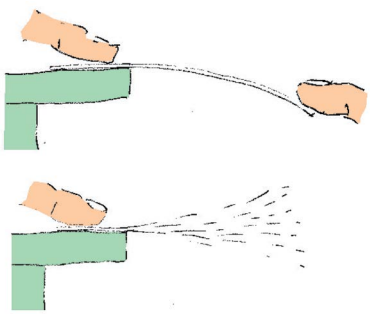

Fig. 1.2 Oscillation of a ruler

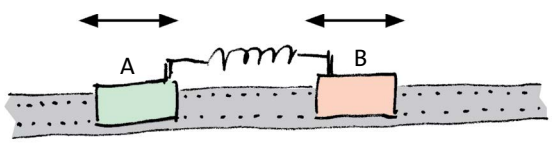

Fig. 1.3 Oscillation of the system „two gliders + one spring“

always the same, regardless of how far the spring is extended at the beginning.

> *Oscillation:*
> - periodic process with a characteristic duration of period;
> - will run by itself after an initial energy supply.

Later, we will not be so accurate anymore: we will slightly attenuate the restrictions of this definition. Even if the process is a bit hampered by friction, we will still call it an oscillation; and even if the period is slightly dependent on the initial energy supply, it will always remain an oscillation.

Eventually, we will even refer to certain processes, which are not even approximately periodic, as oscillations. This matter will be addressed later.

Having formulated the definition in rather general terms comes with an advantage: it will still remain valid if we are dealing with an electric process, i.e. with a process in which not just a body is moving back and forth.

The reciprocal of the oscillation period is called *frequency*. The symbol is *f*.

$$f = \frac{1}{T}. \tag{1.1}$$

The measurement unit 1/second is called Hertz, abbreviated Hz. A short oscillation period is a synonym for a high frequency.

Exercises

1. A driverless wagon that is not slowed down by any friction is located between two spring buffers. It is given a push so that it moves back and forth between the buffers. This process is not called oscillation. Which of the characteristics that an oscillation is supposed to have is missing in the system?
2. Fig. 1.4 shows a surprising experiment that can be performed by means of an construction kit. Two cylindrical shafts are turning opposite to each other at a relatively high angular velocity. A longish bar is laid on the two shafts, i.e. not symmetrically but in a way that its weight puts a higher load on one of the shafts. Then, the bar starts moving back and forth on the two shafts. The process looks like an oscillation. Explain why the bar moves this way. Why isn't the movement an oscillation in the sense of our definition?
3. Fig. 1.5 shows an arrangement that is pretty similar to a steam engine or a combustion engine: a flywheel, a piston rod and a sort of piston that can move back and forth. If the flywheel is given some impetus (i.e. angular momentum), the piston will move regularly back and forth. We disregard losses due to friction. Is the movement of the piston an oscillation according to our definition? Explain.

1.2 Momentum and energy

Now we would like to apply our mechanical know-how to the system from Fig. 1.3.

Let's examine at first how momentum behaves, Fig. 1.6. We release the bodies and they will start moving.

At the beginning, the momentum of A takes on increasingly larger positive values, the momentum of B increasingly larger negative values. The balance is correct: the momentum of A increases by the same amount as that of B decreases. The fact that the spring is under tensional stress tells you that a momentum current is flowing from B to A. But the momentum current will only flow from the right to the left until the spring between the bodies is unstressed. From that moment, the spring will be compressed and the movements of A and B will be slowed down by the spring: momentum will then flow out of A and back into B. Now, the bodies become increasingly slower, eventually come to a halt and will then start moving away from each other again. As soon as the spring is unstressed again, the momentum current will be reversed anew. It will flow once again in the same direction as at the beginning, i.e. from B to A. In its reversal point on the left, A has lost its entire momentum. The system is once again in the same state as at the start, and the next round can begin.

Everything that has been said so far can be summarized in a single phrase: the momentum „slops" back and forth between A and B. Or formulated in a more intellectual fashion:

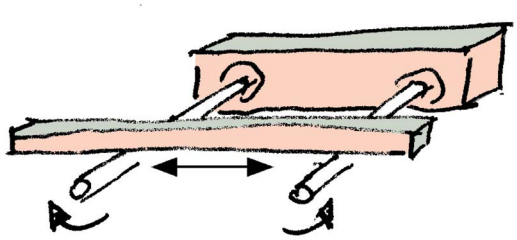

Fig. 1.4 The bar moves back and forth. Is this movement an oscillation? (Exercise 2)

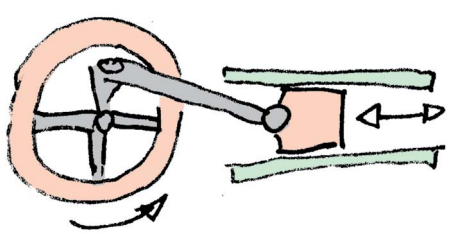

Fig. 1.5 Once boosted, the piston moves back and forth. Is this movement an oscillation? (Exercise 3)

During a mechanical oscillation, momentum flows back and forth between two sub-systems.

Besides, the complicated Figure 1.6 can be simplified, Fig. 1.7. The bent arrow should illustrate the back-and-forth movement of the momentum.

Just as characteristic as the behavior of the momentum is the behavior of the energy for an oscillation.

In the initial state – stretched spring, A and B are at rest – energy is stored in the spring. Its value is calculated according to

$$E_s = \frac{D}{2} s^2.$$

(s is the „deviation" of the spring, D the spring constant.)

If the bodies are released, both of them will start moving. The energy of the spring will decrease, the energy of the two bodies will increase. As we know, the „kinetic" energy of a moving body is calculated according to

$$E_{\text{kin}} = \frac{m}{2} v^2.$$

(v is the velocity of the body, m its mass.)

In the moment in which the spring is unstressed, no energy is stored in the spring anymore. The whole energy is now in the two bodies. As these bodies are approaching each other further, the spring is again charged with energy – at the expense of the two bodies. You can see how things continue.

Also the behavior of the energy can be summarized briefly: it „slops" from the spring outwards into the two bodies, back into the spring etc., Fig. 1.8.

Notice that the slopping back-and-forth of the energy occurs twice as fast as that of the momentum. While the momentum flows back and forth between A and B *once*, the energy moves back and forth between the bodies and the spring *twice*.

During a mechanical oscillation, energy flows back and forth among sub-systems.

The back-and-forth flow of the energy is twice as fast as the back-and-forth flow of the momentum.

Exercise

1. Discussing the oscillation of the system from Fig. 1.3. We have assumed the two bodies to have equal masses. How will the movement change if the masses are different?

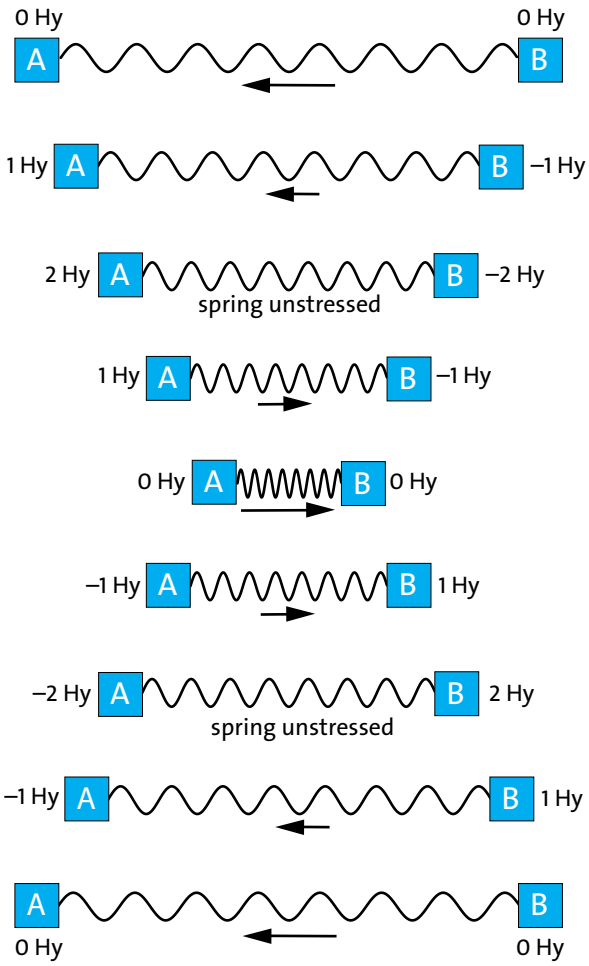

Fig. 1.6 The oscillator from Fig. 1.3 at nine different times in the course of one period. The arrows under the springs illustrate the momentum current.

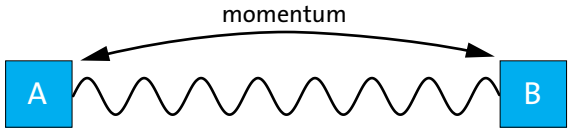

Fig. 1.7 Momentum flows back and forth between the two bodies.

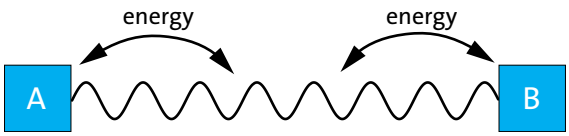

Fig. 1.8 Energy flows back and forth between the bodies and the spring.

1.3 The Earth as a partner

In many cases, the Earth is an essential part of an oscillation system. We can imagine the spring oscillator from Fig. 1.9 to be a result of a modification of the oscillator from Fig. 1.3.

The body is changed with continuously increased mass. In case of Fig. 1.9, A is finally the whole Earth. Now the momentum flows back and forth between body B and the Earth in the course of the oscillation, Fig. 1.10a.

Here, the energy balance is particularly interesting. We might believe that nothing substantial has changed with the energy flow either, i.e. that it would still flow like in Fig. 1.8 with the only difference that A is now the Earth. To understand that this is not correct, we need the equation:

$$E = \frac{p^2}{2m}.$$

We ask for the ratio $E_{\text{Earth}}/E_{\text{B}}$ between the energy contents of the Earth and body B:

$$\frac{E_{\text{Earth}}}{E_{\text{B}}} = \frac{\left(\frac{p_{\text{Earth}}^2}{2m_{\text{Earth}}}\right)}{\left(\frac{p_{\text{B}}^2}{2m_{\text{B}}}\right)} = \frac{p_{\text{Earth}}^2}{p_{\text{B}}^2} \cdot \frac{2m_{\text{B}}}{2m_{\text{Earth}}}.$$

The absolute value of the momentum of the Earth is always equal to the absolute value of the momentum of B. Hence, we have

$$p_{\text{Earth}}^2 = p_{\text{B}}^2$$

and we can reduce the fraction after the right equals sign. We obtain

$$\frac{E_{\text{Earth}}}{E_{\text{B}}} = \frac{m_{\text{B}}}{m_{\text{Earth}}}.$$

Hence, the energy distributes between the Earth and B in the inverse proportion to the masses. But the mass of the Earth is very much larger than that of B. The Earth therefore receives so little energy during the oscillation that it can be safely disregarded. We can consequently say: the energy flows back and forth between B and the spring, Fig. 1.10b. We can also formulate this result in more general terms:

> If one of the two bodies between which the momentum flows back and forth has a much larger mass than the other one, it will no longer take part in the energy turnover of the oscillator.

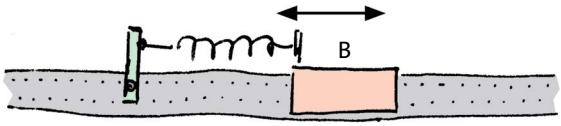

Fig. 1.9 One of the two bodies from Fig. 1.3 is replaced by the Earth.

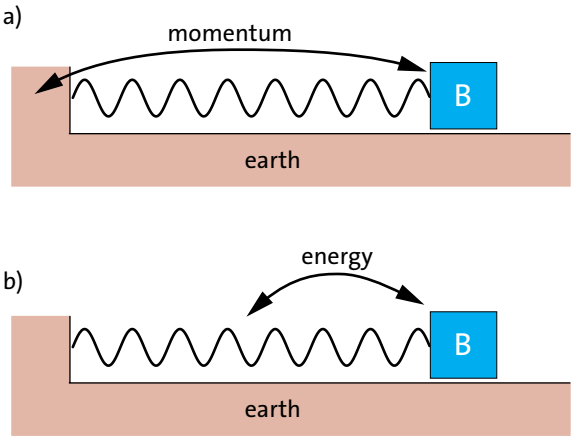

Fig. 1.10 (a) Momentum flows back and forth between body B and the Earth. (b) Energy flows back and forth between body B and the spring with the double frequency.

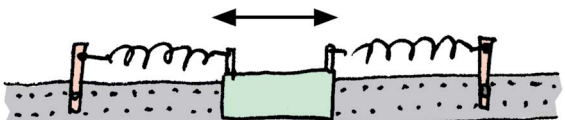

Fig. 1.11 How is the path of the energy and of the momentum?

Exercise

1. Describe the path of momentum and energy when the body from Fig. 1.11 performs oscillations. Check in this process whether the flow direction of the momentum in the springs is correct.

1.4 Harmonic oscillations

We start with some mathematics. For the description of oscillations, a function that you should already know plays an important role: the sine function. For reasons of practicality, it should be written in a specific form:

$$y(t) = \hat{y} \cdot \sin(\omega t + \varphi) \qquad (1.2)$$

($\hat{y}$ is pronounced „wye-hat“)

The corresponding graph is shown in Fig. 1.12. In the following, the time t will always be the independent variable. y stands for the dependent variable in equation (1.2). For y, we will later insert a physical quantity: position, momentum, velocity, electric charge or electric current strength. The equation expresses that the value of the quantity y changes periodically between a maximum value and a minimum value.

Besides t and y, equation (1.2) contains three other symbols: $\hat{y}$, ω and φ. Their meaning can be seen from Fig. 1.13.

$\hat{y}$ is the *amplitude* of the quantity y. It is the maximum value of the function $y(t)$. It is taken on when $\sin(\omega t + \varphi)$ has the value 1. Fig. 1.13a shows two graphs of the function that only differ in the value of the amplitude.

The argument of the sine function

$$\omega t + \varphi$$

is called *phase* at the instant of time t. We would like to understand the meaning of the two constants ω and φ.

If we set $t = 0$, the phase will take on the value φ. Thus, φ is the *starting phase*.

When t is equal to the oscillation period T, the argument of the sine function, i.e. the phase $\omega t + \varphi$, has just increased by 2π. (This is what mathematics teaches us.) Hence, the following must apply:

$$\omega T + \varphi = 2\pi + \varphi.$$

It follows

$$\omega T = 2\pi$$

and

$$\omega = \frac{2\pi}{T}. \tag{1.3}$$

As the frequency is $f = 1/T$, we also have

$$\omega = 2\pi f\,.$$

The *angular frequency* ω in equation (1.2) is – apart from the factor 2π – equal to the frequency and we can also write (1.2):

$$y(t) = \hat{y} \cdot \sin(2\pi f t + \varphi)\,. \tag{1.4}$$

Fig. 1.13b shows two graphs of the function (1.4) that only differ from each other in the value of the frequency f, and Fig. 1.13c shows two graphs that only differ from each other in the starting phase φ.

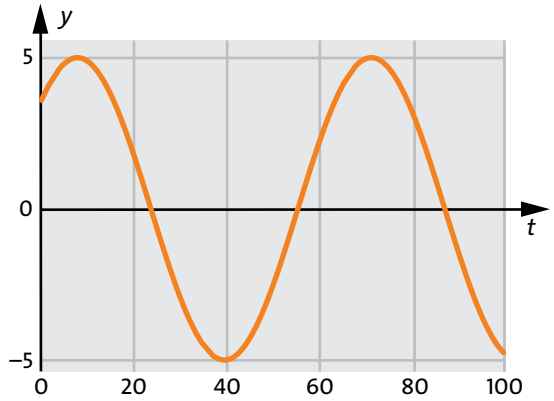

Fig. 1.12 Graph of the sine function. The independent variable is the time t, the dependent variable is y.

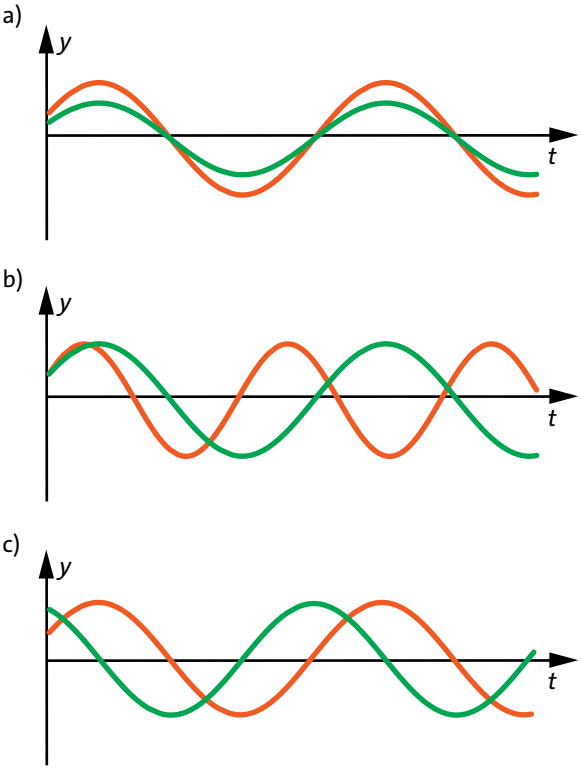

Fig. 1.13 The two sine functions only differ in (a) the amplitude $\hat{y}$, (b) the frequency f and (c) of the starting phase φ.

Now back to physics. When a quantity that is used to describe an oscillation as a function of time is simi-

lar to equation (1.4), the oscillation is called a *harmonic oscillation.* Most oscillations we will examine are harmonic oscillations, but we will see that function (1.4) also plays an important role for non-harmonic oscillations.

An example of a harmonically oscillating system is the arrangement from Fig. 1.3. If there is no friction (or if friction is small that it can be neglected) and if the spring is not overstretched (i.e. if Hooke's law applies), various quantities are described as a function of time by sine functions:

- the momentum of the bodies,
- the momentum current through the spring,
- the position of the bodies,
- the velocities of the bodies

and others.

Also the spring oscillator from Fig. 1.9 performs a harmonic oscillation. We will examine how the different quantities are related to each other in this case.

The position x of body B changes according to a sine function:

$$x(t) = \hat{s} \cdot \sin(2\pi f t) \tag{1.5}$$

$\hat{s}$ is the position amplitude. We have chosen the time zero in a way that the starting phase is $\varphi = 0$. The progression over time of all other quantities we are interested in can be calculated from (1.5).

With

$$v(t) = \frac{d\hat{s}(t)}{dt}$$

by deriving with respect to time we obtain

$$v(t) = \hat{s} \cdot 2\pi f \cdot \cos(2\pi f t) \, .$$

As

$$\cos x = \sin(x + \pi/2)$$

always applies, we can transform this equation in a way that it contains the sine function instead of the cosine function:

$$v(t) = \hat{s} \cdot 2\pi f \cdot \sin(2\pi f t + \pi/2) \, .$$

If we then only rename the term $\hat{s} \cdot 2\pi f$ to $\hat{v}$, we obtain an expression that has once again the structure of equation (1.2):

$$v(t) = \hat{v} \cdot \sin(2\pi f t + \pi/2). \tag{1.6}$$

Here, $\hat{v}$ is the velocity amplitude.

In contrast to the position function (1.5), the starting phase is no longer zero in (1.6). The phase of the function $v(t)$ is at all times greater than that of $s(t)$ by $\pi/2$. We also say that there is a *phase difference* of $\pi/2$ between $v(t)$ and $s(t)$.

Also the momentum as a function of time can be calculated from (1.6). With

$$p(t) = m \cdot v(t)$$

we get

$$p(t) = m \cdot \hat{v} \cdot \sin(2\pi f t + \pi/2) \, .$$

If we denominate the momentum amplitude $m \cdot \hat{v}$ with $\hat{p}$, we obtain:

$$p(t) = \hat{p} \cdot \sin(2\pi f t + \pi/2), \tag{1.7}$$

hence, a sine function once again.

The energy as a function of time in the two energy storage devices „body B" and „spring" is more interesting. We will calculate both energy contents.

We start with the spring. Its energy content can be calculated according to

$$E_S = \frac{D}{2} s^2.$$

We insert (1.5):

$$E_S(t) = \frac{D}{2} \cdot \hat{s}^2 \cdot \sin^2(2\pi f).$$

Also this function term contains the sine function. But here it is squared. In Fig. 1.14a, the spring energy is illustrated as a function of time.

The fact that the square of the sine function is itself a sine function, which, however, is displaced in the direction of the ordinate axis and which oscillates at twice the frequency can be concluded from a generally applicable equation that you might have learned earlier in math class

$$\sin^2\alpha = \frac{1}{2} - \frac{1}{2}\cos(2\alpha).$$

Besides the energy $E_S(t)$ of the spring, also its deviation $s(t)$ is shown in Fig. 1.14a.

A comparison of the two functions shows:

The curve belonging to the energy is sine-shaped. However, the sine function is displaced in the direction of the ordinate axis so that the energy values will never become negative – which would not be possible,

of course. The lowest energy value is 0 J. The frequency of the energy sine function is twice as high as the frequency of the *s*-sine function.

We now calculate the energy content of the body in the same way:

$$E_{\text{kin}} = \frac{p^2}{2m}.$$

We rewrite (1.7):

$$p(t) = \hat{p} \cdot \sin(2\pi f t + \pi/2) = -\hat{p} \cdot \cos(2\pi f t)$$

and insert

$$E_{\text{kin}}(t) = \frac{\hat{p}^2}{2m} \cdot \cos^2(2\pi f t).$$

The function graph is shown in Fig. 1.14b. It looks like that of $E_S(t)$ with the exception of being displaced with respect to $E_S(t)$ in such a way that its maximum values are located at the same points as the minimum values of $E_S(t)$ and vice versa. The sum of the two functions

$$E_S(t) + E_{\text{kin}}(t)$$

is constant, its function graph is a horizontal straight line, Fig. 1.14c. This simply means that the total energy content of the system does not change. The energy flows back and forth between the spring and the body but the total amount of energy does not change.

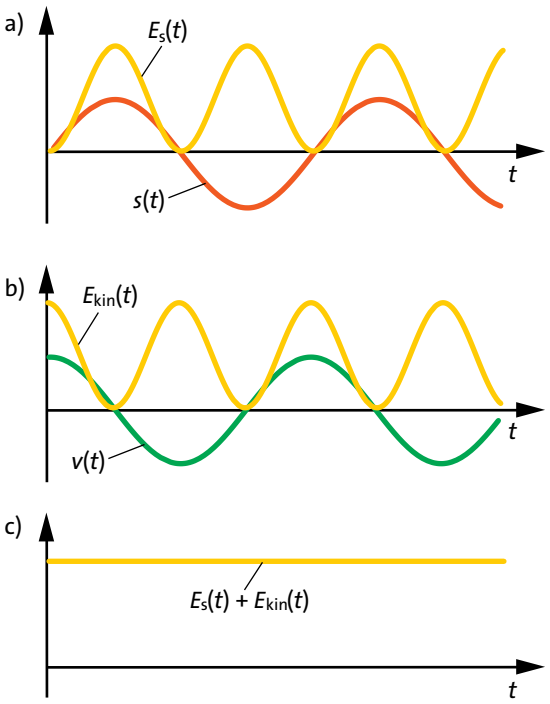

Fig. 1.14 (a) Deviation of the spring and energy of the spring as a function of time. (b) Velocity and kinetic energy of the body as a function of time. (c) The sum of the energy of the spring and the energy of the body is constant.

Exercises

1. Calculate for the spring oscillator from Fig. 1.9 the function term for the momentum current that flows back and forth between the body and the Earth. Graphically illustrate the momentum of the body and the momentum current between the body and the Earth in a diagram. Discuss the relationship between the two function graphs.
2. Calculate for the spring oscillator from Fig. 1.9 the function term for the energy current that flows back and forth between the body and the spring. The result contains a combination of trigonometric functions. Try to simplify this term. You will probably need a mathematical formulary. Discuss the result in connection with Fig. 1.14.
3. The position of body A in Fig. 1.3 was described by the function $s = \hat{s} \cdot \sin(2\pi f t)$. What is the corresponding function for body B? Indicate the function terms for: the momentum of the two bodies; the momentum current between A and B; the energy of the spring and the two bodies individually; the total energy; the energy current between the bodies and the spring.

1.5 What the period length depends upon

We will examine the question on the example of the oscillator from Fig. 1.9 because it is the simplest system for this purpose. We already know that the oscillation period is independent of the amplitudes of position, velocity or momentum. We now would like to make the oscillator oscillate more slowly. What could be done? There are two possibilities.

1. Changing the oscillating body

We try it out. We double the mass by attaching a second glider to the first one. The oscillation becomes significantly slower. It would not be unreasonable to expect the oscillation period to double. We check by measuring and find out that our expectation is not fulfilled. The oscillation period has indeed increased, but to less than the twice its initial value. We add another body and measure once again. Also this time, the period length increases but it still has not reached the double value. Only when four bodies are connected to the spring, i.e. after having quadrupled the mass, the

period will double, Fig. 1.15. By how much would the mass have to be increased in order to achieve a triple period?

We conclude: for the oscillator from Fig. 1.9, the oscillation period is proportional to the square root of the mass of the oscillating body.

2. Changing the spring

If it is replaced by a softer spring, i.e. by a spring with a smaller spring constant, the oscillation will become slower, i.e. the oscillation period will become longer. We proceed in a similar way as during the examination of the dependence on the mass. We change the spring constant to half its initial value. This is not difficult either: we simply add another spring to the first spring. (In analogy to electric resistors we could say: „two springs are connected in series".) This time, the result is no longer surprising: the period has increased but not doubled. To reach a double value, we have to combine four springs, Fig. 1.16.

Again, we conclude:

For the oscillator from Fig. 1.9, the period is proportional to the reciprocal value of the root of the spring constant.

Both results can be summarized to one relationship:

$$T \sim \sqrt{\frac{m}{D}}.$$

To make an equation of the proportionality, it is sufficient to measure the period of an oscillator with any mass and spring constant. We find that the factor 6.3 has to be ahead of the square root. Addressing the problem theoretically would lead us to the more precise equation:

$$T = 2\pi\sqrt{\frac{m}{D}}. \qquad (1.8)$$

By means of equation (1.3), we obtain the angular frequency:

$$\omega = \sqrt{\frac{D}{m}}.$$

This equation might appear complicated to you. In addition, the corresponding formula looks different for other systems. Can we memorize something like this? When you have more experience with other oscillation systems, you will notice that the frequency formulas (or oscillation period formulas) can be obtained by means of skilled guessing. At first, you reflect on the quantities the angular frequency must depend on. In the simple examples

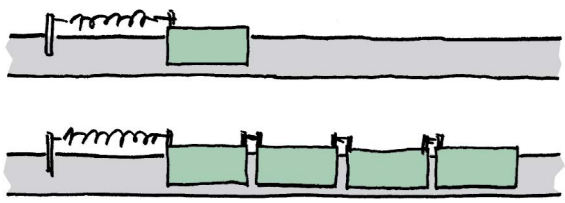

Fig. 1.15 The mass is increased by adding further, identical bodies. Quadrupling the mass leads to doubling of the oscillation period.

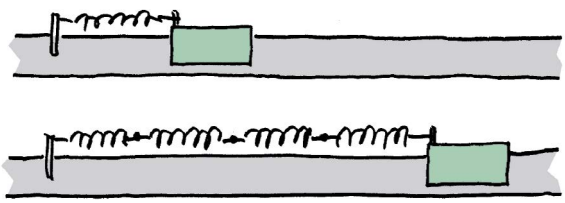

Fig. 1.16 The spring constant is decreased by attaching several identical springs to one another. If the spring constant is reduced to a quarter, the oscillation period will double.

that we examine, it always depends on two quantities. In the previous example, these quantities were the mass of the body and the spring constant of the spring. Next, you consider whether the angular frequency increases or decreases with the quantities. If it increases with a quantity (here with the spring constant), this quantity will stand in the numerator; if it decreases (in our case with the mass), this quantity will stand in the denominator. Then, only the square root of the fraction obtained this way has to be calculated.

Exercises

1. An oscillator like the one in Fig. 1.9 has a period of 2 s. The mass of the body is 250 g. Which mass would be needed to increase the period to (a) 3 s, (b) 10 s. (The spring constant should be kept constant.)
2. How will the oscillation period behave if a further, identical spring is added to the first one in an oscillator of the type from Fig. 1.9? What will happen if a total of four identical springs are arranged in parallel?
3. The two bodies in Fig. 1.3 have the masses m, the spring has the spring constant D. What is the formula for the period length? Clue: divide the system into two parts to which the formula (1.8) can be applied.
4. The body in Fig. 1.11 has the mass m, each of the springs has the spring constants D. What is the formula for the oscillation period?

1.6 The pendulum

An object that is suspended on a thread or a rope can perform an oscillation movement. This process is probably the best-known oscillation process, but not the simplest one.

Such an arrangement that has been built deliberately for the purpose of performing oscillations is called *pendulum*. Instead of a flexible thread, also a rigid bar is often used for suspension. In this case, the pendulum can oscillate even beyond the horizontal position.

At first, we simply observe the movement of a pendulum. The angle α by which the pendulum deviates from the vertical resting position is suitable as a quantity to characterize the position of the pendulum, Fig. 1.17.

Figure 1.18 shows this angle as a function of time for different initial deviations. Be aware that the scale of the ordinate axis is different in the three subframes. A comparison of the three images shows:

1. The oscillation is not harmonic for large amplitudes. For sufficiently small amplitudes it becomes approximately harmonic.

2. For large amplitudes, the oscillation period depends on the amplitude. For sufficiently small amplitudes it becomes practically independent of the amplitude.

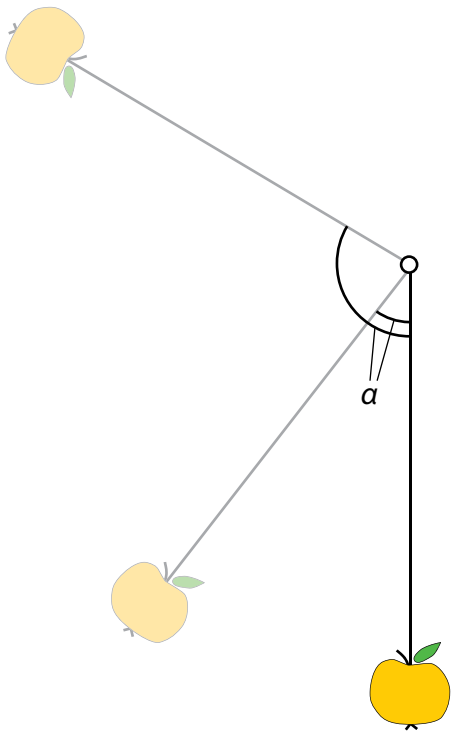

Fig. 1.17 The angle α is a measure for the deviation of the pendulum from its position of rest.

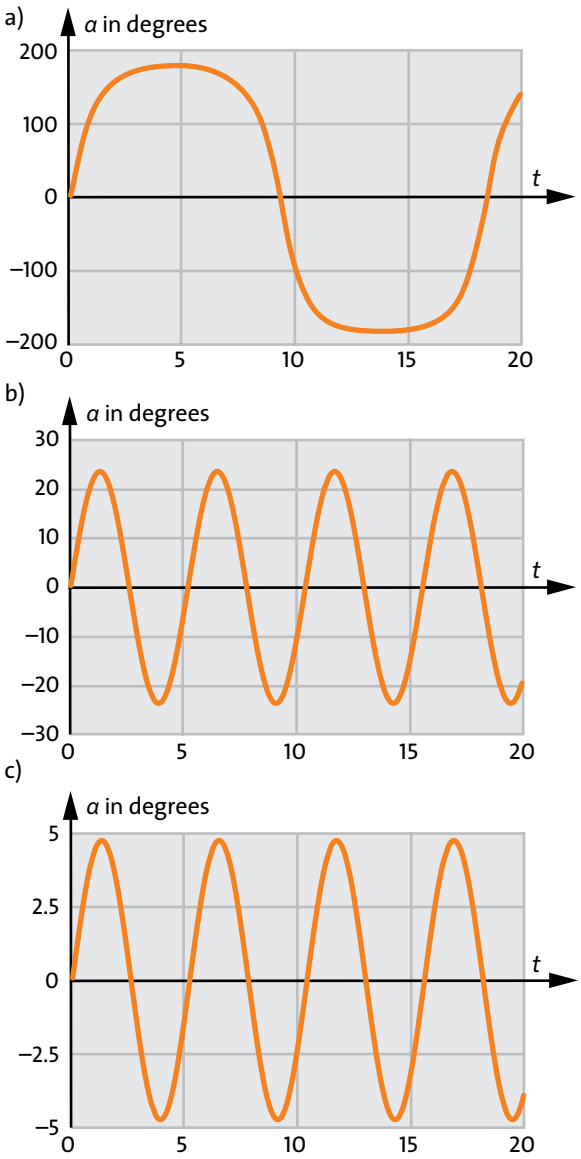

Fig. 1.18 The angle α as a function of the time t. (a). Oscillation with a very large amplitude, shortly prior to overshooting. The function graph is not a sine curve. (b) The amplitude of $\alpha(t)$ is now only 23°, i.e. much less than in (a). The function graph is now a sine curve. The period length is significantly smaller than in (a). (c) The amplitude of $\alpha(t)$ is slightly less than 5°, i.e. much less than in (b). The function graph is still a sine curve, and also the period length has not changed anymore.

The momentum balance of the pendulum

It is more complicated than in the systems examined earlier because the movement takes place in two dimensions. If, however, we limit our experiment to small initial deviations, i.e. to the case that the oscillation is harmonic, also the momentum balance will remain quite simple. The vertical component of the momentum vector of the body will then always be much smaller than the horizontal component and it makes sense to limit the examination to the horizontal component. This horizontal momentum changes approximately as a sine function. It flows back and forth between the oscillating body and the Earth through the rope and the suspension of the pendulum. It is not difficult to perform the experiment shown in Fig. 1.19. While the body swings back and forth, also the vehicle with the suspension moves back and forth, always in the opposite direction. We can therefore clearly see where the horizontal momentum is located in every moment.

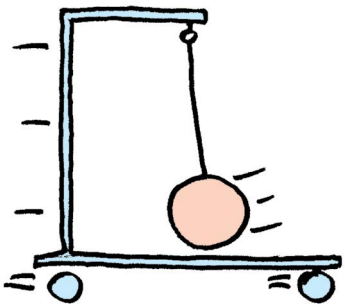

Fig. 1.19 The momentum oscillates back and forth between the swinging body and the vehicle.

The energy balance of the pendulum

If the pendulum is deviated, the body will necessarily be slightly lifted. Therefore, energy is supplied to the gravitational field. If the pendulum is then released, the body will start moving while its height coordinate decreases. Thus energy from the gravitational field comes back and the kinetic energy of the body increases. After having passed the deepest point, the body becomes slower again. Energy flows out of the body and back into the gravitational field. The formulas for the energy content of the two storage systems „gravitational field“ and „body“ are:

$$E_g = m \cdot g \cdot h$$

and

$$E_{kin} = \frac{m}{2} v^2.$$

In each oscillation period, the energy flows back and forth twice between these two storage devices – similar as in the case of the spring oscillator.

The oscillation period of the pendulum

It would be best if you could try to find out the formula yourself.

We do not make a detailed examination at the start. The period probably depends on the mass of the body, possibly also on the pendulum length. We try it out. First, we measure the period of two pendulums that only differ from each other in the mass of the oscillating body. The thread length should be equal. The measurement shows a surprising result: the oscillation period is identical for both pendulums. Hence, we can make the pendulum as heavy or light as we wish – the period remains the same. Apart from this, we examine whether the period depends at least on the length of the thread. This time we have more success. It turns out that the period is proportional to the square root of the length of the pendulum. But should the period only depend on the length of the pendulum? You certainly have a good intuition for mechanical processes and you will suggest to repeat the measurement on the Moon. In the next physics class, you can certainly go on a short trip up there. You would find: the pendulum oscillates more slowly. Apparently, the period depends on the gravitational field strength g. The measurement shows that the dependence follows a square root law again.

The definite formula is:

$$T = 2\pi\sqrt{\frac{l}{g}}.$$

The similarity with equation (1.8) is noticeable.

The fact that the period is independent of the amplitude (at least as long as the amplitude is small) was taken advantage of to build the old pendulum clocks, Fig. 1.20. The clock hand advances by a small step with each oscillation period of the pendulum.

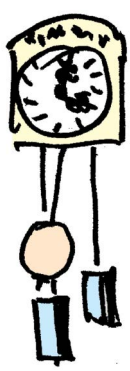

Fig. 1.20 Pendulum clock

The two weights, i.e. one for the clockwork mechanism and the other one for the bell, are used as energy sources of the clock. Each weight is suspended on a cord that is coiled up on a drum while the clock is being wound up. This process is used to store the energy in the gravitational field. While the clock is running, the weights are lowering slowly. Energy thereby comes back from the gravitational field and goes to the clockwork mechanism and the bell via the drums.

The progression of the clock hands is controlled by the pendulum by means of a sophisticated mechanism: after every half oscillation of the pendulum, the minute hand makes a tiny step forward (and also the hour hand via a 12 : 1 gearwheel transmission system).

All modern clocks also work according to this principle. But other oscillation systems are used instead of a pendulum these days.

Exercises

1. The thread of a pendulum has a length of 1.2 m. What is the pendulum's period? What would be its period on the Moon? Although the pendulum would not survive a journey to a neutron star: Theoretically, what would be the period at the surface of such a star?
2. What must be the length of the pendulum of a clock, so that it swings forth in one second and back again in one second?
3. Which are the disadvantages of a pendulum clock compared with other types of mechanical clocks?
4. Somebody has taken a pendulum clock to the Moon. Does the clock work there? (Which problems arise for the pendulum, and which for the weights?)
5. A body with a mass of 2 kg is suspended on a string with a length of 2 m. With a short stroke the body is brought to a velocity of 0.2 m/s. Which height will the body reach? (Attention: The question if only for the height.) The same body now hangs an a string with a length of 10 m. What height does it reach now?

1.7 Angular oscillations: angular momentum flowing back and forth

A flywheel is connected to the „Earth" via a spiral spring, Fig. 1.21. The arrangement is a system that can perform harmonic oscillations.

If the flywheel is turned out of its equilibrium position and released, it will rotate back and forth in a sine-shaped way. The similarity to the oscillator from Fig. 1.9 is obvious. Everything that has been said in connection with this oscillator also applies to the angular oscillator provided that we translate the respective quantities as follows:

- position x → angle α
- velocity v → angular velocity ω
- momentum p → angular momentum L

Here, angular momentum flows back and forth between the flywheel and the Earth while energy flows between the flywheel and the spring, Fig. 1.22. Compare with Fig 1.10.

Also the formula for the oscillation period is similar to the respective formula for the oscillator from Fig. 1.9. However, it includes quantities that you have not yet learned about. It is not worth deriving the details of this formula. We would only like to examine what the period depends on after all.

A sort of spring constant can also be introduced for spiral springs so that the „hardness" of the spring can be characterized. The harder the spiral spring, the faster the oscillation and hence the shorter the period.

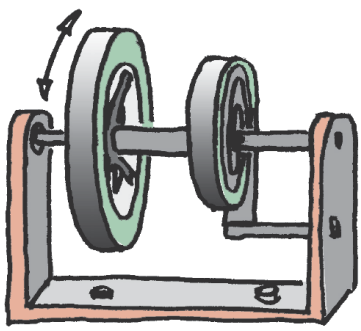

Fig. 1.21 The system that consists of a flywheel, a spiral spring and the Earth performs angular oscillations.

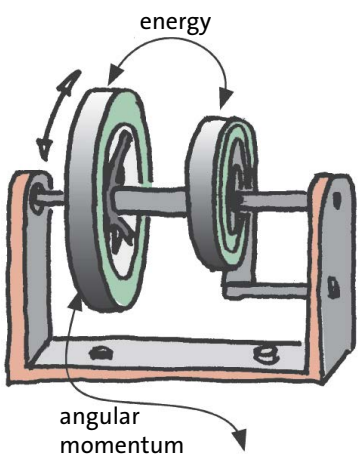

Fig. 1.22 Angular momentum flows back and forth between the flywheel and the Earth. The energy flows back and forth between the flywheel and the spring at the double frequency.

It is logical and easy to verify that the period will become longer if the mass of the flywheel is increased.

The fact that the period length also depends on where the masses are sitting makes this matter slightly more complicated than a normal back-and-forth oscillation. If the flywheel has a large radius so that the masses are sitting far towards the outside, it will oscillate more slowly than in cases where the masses are concentrated at the center. This dependence can be examined most conveniently if a sort of dumbbell is made oscillate instead of a flywheel, Fig. 1.23. Here, the distance of the two bodies from the center can be increased and reduced without changing the total mass. We find that the period length is proportional to this distance.

The angular oscillator has been built in hundreds of millions of copies. Named „balance wheel", it is used to control the hands of all mechanical wrist watches, pocket watches and alarm clocks; see exercise 2.

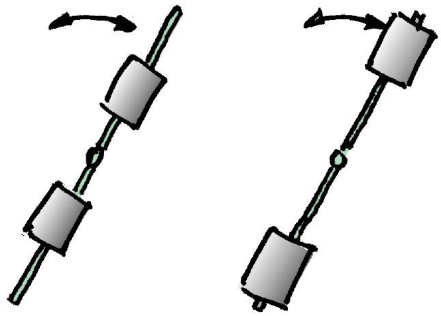

Fig. 1.23 Also a „dumbbell" can be used instead of the flywheel. The farther to the outside the two bodies are sitting, the slower the oscillation of the dumbbell.

Exercises

1. An angular oscillator can also be built in analogy to the system from Fig. 1.3. What does it look like? How do the angular momentum and the energy flow?
2. What are the advantages of the balance wheel, in contrast to a pendulum and to an oscillator with a helical spring as shown in Fig. 1.9, as a timer of a clock?
3. Carefully open a mechanical wrist or pocket watch or a mechanical alarm clock and look for the balance wheel. Try to understand how the balance wheel controls the clockwork. Describe.

1.8 Electric oscillations: electric charge flowing back and forth

It would be best for you to close this book and try to invent an electric oscillator yourself.

The simplest solution is shown in Fig. 1.24. The system is very similar to the mechanical oscillator from Fig. 1.3 that we show here once again as Fig. 1.25.

Just as the momentum „swashes back and forth" between the two bodies A and B in the case of a mechanical oscillator, the electric charge oscillates back and forth between the plates A and B of the capacitor in the case of the electric oscillator. In the connection between the plates A and B of the capacitor a coil has to be installed. It is analogous to the spring in Fig. 1.25. (But the fact that the illustrations of a coil and a spring look similar is a pure coincidence.) The capacitor has

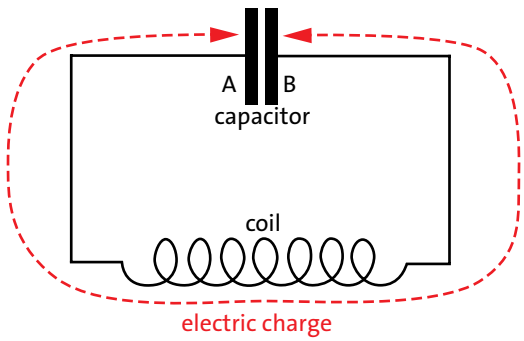

Fig. 1.24 Resonant circuit. The electric charge flows back and forth between the two capacitor plates.

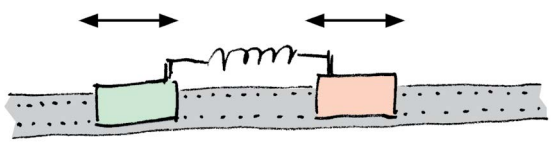

Fig. 1.25 Mechanical oscillator, similar to Fig. 1.3

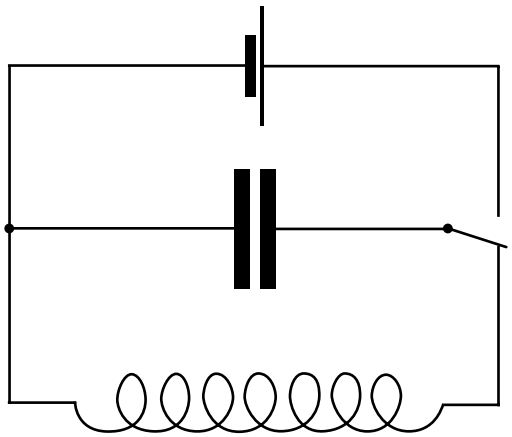

Fig. 1.26 The switch is turned upwards to charge the capacitor. It is turned downwards to start the oscillations.

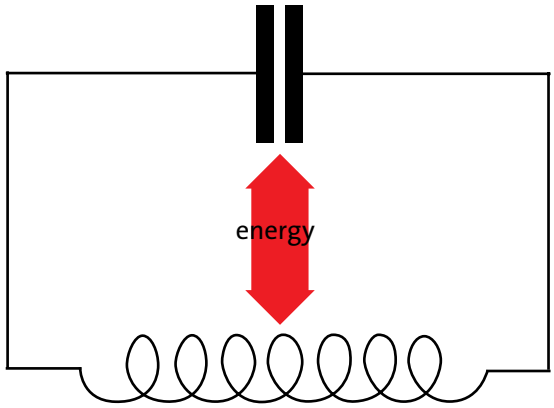

Fig. 1.27 The energy flows back and forth between the capacitor and the coil.

to be charged at first in order to enable the oscillation to start. Fig. 1.26 shows how this is done in practice.

In a resonant circuit, the electric charge flows back and forth between the plates of a capacitor.

As the electric oscillator has the structure of a circuit, the arrangement from Fig. 1.24 is also called *resonant circuit*. Just like the mechanical systems examined earlier, the resonant circuit performs harmonic oscillations. The electric charge on one of the capacitor plates changes over time according to a sine function:

$$Q(t) = \hat{Q} \cdot \sin(2\pi f t) \qquad (1.9)$$

(The charge of the other plate only differs in its sign.)

The change of the electric charge of the plates goes together with an electric current. When the change dQ/dt is large, a strong electric current flows. This fact is expressed by the equation

$$I(t) = \frac{dQ(t)}{dt}.$$

Thus, the electric current strength in the conductor between the plates of the capacitor is obtained by deriving the function $Q(t)$ with respect to time:

$$I(t) = \hat{Q} \cdot 2\pi f \cdot \sin(2\pi f t + \pi/2).$$

We abbreviate the current amplitude $\hat{Q} \cdot 2\pi f$ as $\hat{I}$ and obtain:

$$I(t) = \hat{I} \cdot \sin(2\pi f t + \pi/2).$$

The comparison with (1.9) shows that the current strength is „phase-shifted“ by $\pi/2$ with regard to the charge.

Just as the energy in the case of the mechanical oscillator flows back and forth between the body and the spring at twice the frequency, it flows back and forth between the capacitor and the coil at twice the frequency in the resonant circuit – more precisely: between the electric field in the capacitor and the magnetic field in the coil, Fig. 1.27. When the charge of the capacitor reaches its maximum, the whole energy is located in the capacitor and there is no energy in the coil. When the charge of the capacitor is zero, the whole energy is in the coil.

In a resonant circuit, energy flows back and forth between a capacitor and a coil.

The back-and-forth flow of the energy is twice as fast as the back-and-forth flow of the electric charge.

The period can be calculated from the capacitance C of the capacitor and the inductance L of the coil:

$$T = 2\pi\sqrt{L \cdot C}\,.$$

Just as in the corresponding formulas for mechanical oscillations, there is the square root and the factor 2π.

Resonant circuits are not only a nice physical toy. They have a high relevance for technical applications. To transmit messages by means of electromagnetic waves, a sender and a receiver is needed. A resonant circuit has to be installed in both of them. The one in the sender is used (together with the transmitting antenna) to create the wave. The one in the receiver captures the signal

with the desired frequency out of a multitude of other signals that come from the receiver antenna.

Exercises

1. Which geometrical quantities do the capacitance and the inductance depend on? What is the structure that the capacitor and the coil need to have in order to enable a long period of the respective resonant circuit?
2. What is the time function of the energy that is contained in the electric field of the capacitor? What is the time function of the energy in the magnetic field of the coil?

Fig. 1.28 Swing door

1.9 The damping of oscillations

Up to now, we have regarded friction, that causes an oscillation to decay or die away, as a disturbance that should be avoided in the best possible way. In fact, however, there are systems in which, although they can oscillate in principle, the oscillation is undesired. A swing door, Fig. 1.28, has a spring that is used to make the door shut by itself after someone has passed it. However, some doors oscillate back and forth several times before they stop. Hence, this might cause some inconvenience for the next person wishing to pass through said door. In this case, it would therefore be useful to provide a *damping* system for the oscillation.

Fig. 1.29 schematically displays a system that performs *damped oscillations*. It differs from the system in Fig. 1.9 due to the damper that is installed in parallel to the spring.

Fig. 1.30 shows the design of a damper. When the piston is moved in one or the other direction, the liquid has to flow through a small hole from one to the other side. The faster the piston is moved, the harder it becomes or, formulated with physical quantities, the higher the velocity of the piston relative to the cylinder of the damper, the stronger the momentum current through the damper. Hence, the bars with the corresponding loops on both sides of the damper are the inlet and outlet for a momentum current.

We can therefore describe the effect of the damper as follows:

> The greater the velocity difference between the inlet and the outlet, the stronger the momentum current through the damper.

In the simplest case, the relation between the velocity difference Δv and the momentum current strength F_D is linear so that the following applies:

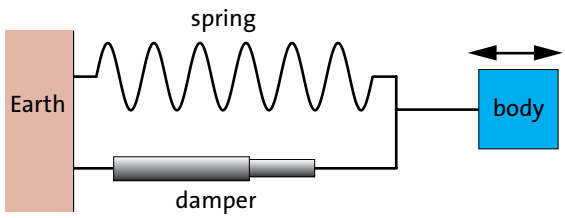

Fig. 1.29 Oscillator with a damper, schematic display

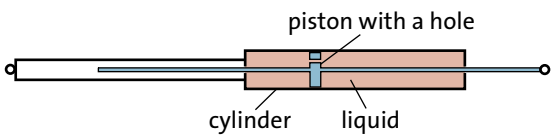

Fig. 1.30 Mechanical damper

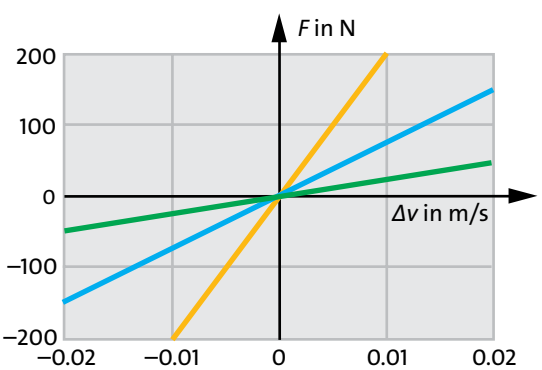

Fig. 1.31 Relationship between momentum current strength and velocity difference for three different dampers

$$F_D = k \cdot \Delta v$$

Fig. 1.31 shows the relation for three different dampers. The higher the value of the damping constant k, the steeper the straight line and the harder the damper. Or in other words: a damper that is hard to compress is a good conductor for a momentum current; a damper that is easy to compress is a bad momentum current conductor.

Even if we do not install a damper, i.e. if there is only the inevitable natural friction, we can often assume terms that the respective momentum current approximately depends linearly on the velocity.

Maybe you have noticed that a damper is the same for a momentum current as an electric resistor is for an electric current, because the following applies for an electric resistor:

The greater the electric potential difference between the inlet and the outlet, the stronger the electric current through the resistor.

When the electric voltage is proportional to the electric current strength, we say that Ohm's law applies.

We come back to our swing door that should be equipped with a damper. What should the damper be like in order to make the door convenient to use? Should it rather be soft or hard? If it is too soft, it will only slow down the oscillation slightly. Although the door will oscillate a bit less, it will not stop completely. So should we rather choose a very hard damper? If we do, the door will no longer oscillate but it will have another defect: it will move very slowly so that it will take long to shut completely.

Fig. 1.32 shows the movement of an oscillator for 5 differently chosen dampers. In our case, the coordinate y would stand for the angle by which the door is moved out of its closed position.

In the first case, the damping effect is weak. The door oscillates back and forth for a long time. In the second case, the damping effect is a bit stronger, i.e. the door comes to a resting state faster. In the third case, it shuts very quickly. It remains shut after a single overshoot. In the fourth picture, an even harder damper was chosen. But the door does not shut faster now. Although it does not oscillate any longer, it needs more time to shut. In the last image with an even harder damper, the situation is even worse. It takes very long until the door will be shut completely. We can see that there is an optimal damping case if we want the oscillator to stop moving after a short time.

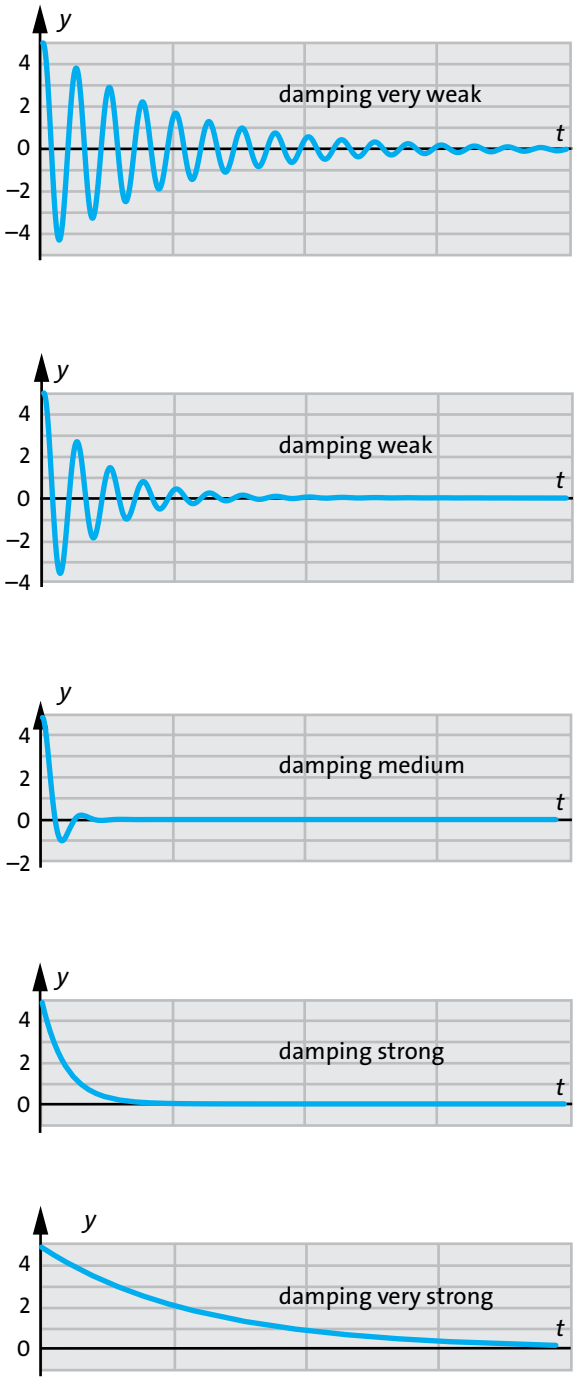

Fig. 1.32 Decay behavior of an oscillator for 5 different damping systems. The dampers become increasingly harder from the top towards the bottom. For a very weak damping effect, the system needs a long time to gradually stop oscillating. For a very strong damping effect, it reaches the equilibrium state very slowly. In case of an optimally chosen damping mechanism, it returns to the equilibrium situation very quickly.

In the case of an optimal damper, an oscillation system will return to the state of equilibrium with a minimum time.

We have discussed the usefulness of a damper with the example of the swing door. But there are countless other situations in which we look for a fast-decaying oscillation.

The needles of any meters should reach their final position as fast as possible. They should not oscillate around the measurement value for a long time, and neither should they approach the definite measurement position too slowly. Therefore, meters with needles are always equipped with an appropriate damper.

The most important oscillation damper application is probably in vehicles: in passenger cars and trucks, locomotives and railway cars, motorcycles, mopeds and even in certain bicycles.

In the following, we will sometimes need to make reference to the different masses of a vehicle. We name them in the way that is illustrated in Fig. 1.33: the part of the vehicle that is carried by springs shall be called H (for „heavy“), the part between the springs and the road (or rail), i.e. the axes, wheel suspensions, the wheels and any associated parts in case of driven wheels, shall be referred to as L (for „light“).

Vehicles need a damping system because they have springs. What are the springs used for? You can certainly imagine how we would feel in a vehicle without any springs. The movement of H would follow the irregularities of the routes, the roads or the rails. The springs ensures the movement of H to be smoothened.

Now we imagine a car that has a spring system but no damper yet. What will happen if such a car drives on a road that is not completely even? The car will make movements that are not quite as we want them to be in two respects.

1. Together with the 4 springs, part H of the vehicle forms a system that is essentially identical to our spring oscillator from Fig. 1.9. H is nudged by every irregularity of the road so that the car begins to performs an oscillation in the vertical direction and also keeps swaying for some time after passing the irregularity. This is not convenient for the passengers.

2. After driving over an irregularity, a wheel makes a bouncing movement similar to a rubber ball. During such a bouncing movement, the wheel has no contact to the road anymore for most of the time. With a wheel that is currently in the air, however, neither braking nor accelerating or steering is possible. It would consequently be dangerous to drive such a car.

Both the oscillation movement of the vehicle part H as well as the bouncing movement of the vehicle part L

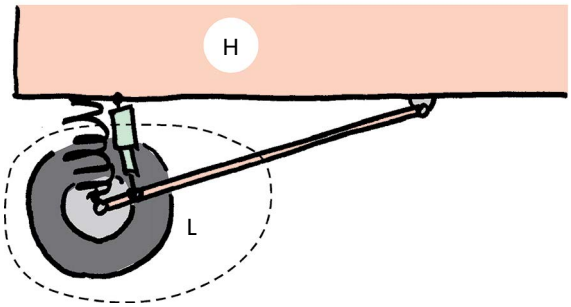

Fig. 1.33 The heavy part H of the vehicle is connected to the light parts L via the springs and the dampers.

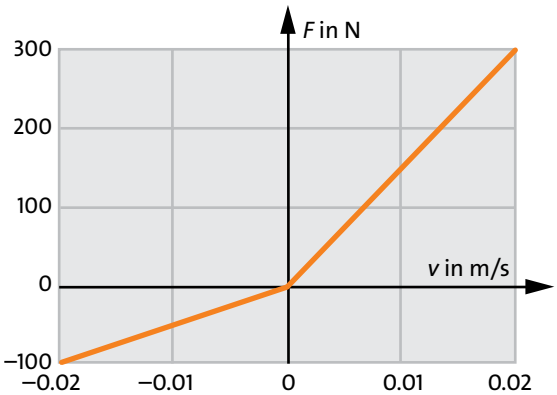

Fig. 1.34 Relationship between momentum current strength and velocity difference for the shock absorber of a car

is now damped away by means of the so-called shock absorbers – one for each wheel.

The shock absorber of a car essentially has the structure shown in Fig. 1.30 and is installed „in parallel“ to the spring; see Fig. 1.33. Also, the spring and the shock absorber sometimes form a single module.

A subtle difference between a real car shock absorber and the dampers that we examined earlier becomes clear in Fig. 1.34. The shock absorber is asymmetric: during compression, it is softer than while being pulled apart. Can you imagine why this is done?

Do you also understand now why the shock absorbers are checked during the legally required technical inspections?

Exercises

1. Identify for a car, a truck, a railway-car and a motorcycle the springs and the shock absorbers.
2. The oscillator of Fig. 1.3 is to be damped. Make a schematic drawing that shows how the damper has to be built in. Sketch the flow of momentum and energy in the way of Fig. 1.7 and Fig. 1.8.

2 RESONANCE

2.1 Resonance

Suspend a heavy object (several kg of weight) on a cord and hold the cord on the other end. Then, move the upper end of this pendulum back and forth slowly, Fig. 2.1a. The object follows your movement but does not quite do what we would call an oscillation. Now accelerate the back-and-forth movement of your hand gradually. The pendulum starts to oscillate, the thread will no longer remain vertical. If you increase the frequency of the back-and-forth movement further, the oscillation of the pendulum will become increasingly intense. At even higher frequencies, however, it will slow down again. It will simply not manage to keep pace.

Hence, at a well-defined frequency of the movement of your hand, the pendulum reacts strongest. This phenomenon is called *resonance*.

Maybe you have noticed that you have to make more or less of an effort to perform the back-and-forth movement. At first, i.e. at a lower frequency, less and then, during resonance, a lot, and finally – i.e. at high frequencies – less again. In case of the resonance, we apparently „pump" much energy into the oscillator. We can also conclude from another fact that things have to be like this: during resonance, the pendulum moves fastest. Therefore, friction is strongest and a maximum of entropy is created. For the creation of entropy, however, energy is required.

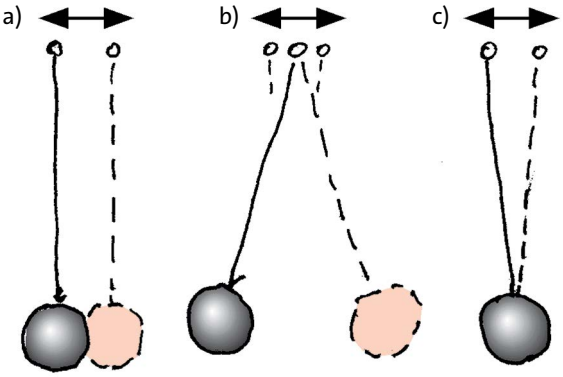

Fig. 2.1 „Excitation" of a pendulum (a) with a low frequency (b) with the natural frequency (c) with a high frequency

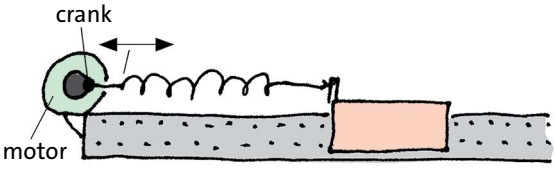

Fig. 2.2 The spring oscillator is excited by means of an electric motor.

2.2 Resonance of a mechanical oscillator

Although the pendulum is particularly easy to set up, its theoretical description is a bit complicated. We therefore come back to our spring oscillator. It can also be used to observe the phenomenon of resonance: we move the left end of the spring sinusoidally (like a sine function) back and forth, Fig. 2.2. For this purpose, we use a motor (i.e. a source of energy) with a crank handle. The fact that the spring does not only move back and forth but also up and down in this process will not disturb the experiment.

Fig. 2.3 is a schematic illustration of the oscillator. As friction will be important in the following, we have illustrated it by means of a damper. The damper symbol stands for both the natural, i.e. involuntary, friction as well as for real damper that might exist.

First, some technical terms have to be introduced.

The frequency at which the spring oscillator (or any other oscillator) oscillates after being nudged once and then let up to itself is called *natural frequency* f_0 of the oscillator. The motor with the crank handle is called exciter. The frequency of the sine-shaped back-andforth movement of the motor is called *exciter frequency.*

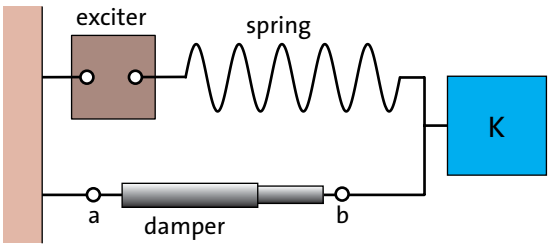

Fig. 2.3 Schematic display of a mechanical oscillator with a damper and an exciter

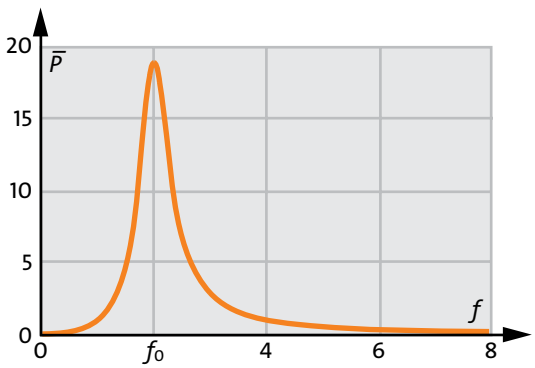

Fig. 2.4 Resonance curve: the average energy current from the exciter to the oscillator as a function of the frequency

Now we examine the movement of the oscillator as a function of the exciter frequency. We find out the following:

1. Body K makes a harmonic movement regardless of the value of the exciter frequency f. Would you have expected that? It means that the time dependence of the position, the velocity, the momentum and other quantities is given by a function of the type:

$y(t) = \hat{y} \cdot \sin(2\pi f t + \varphi)$

The values of the amplitudes of these quantities, however, depend on the frequency of the exciter.

2. We can see best how the oscillator reacts to the excitation if we ask how much energy it „absorbs“ per time. We mean the energy that flows from the exciter to the oscillator and that is used in the exciter for the production of entropy (due to mechanical friction). As this energy current changes constantly in the course of an oscillation period, we look for its time average value $\overline{P}$.

Hence: we are interested in the average energy current from the exciter to the damper as a function of the exciter frequency f, i.e. for the function $\overline{P}(t)$. Fig. 2.4 shows the function graph. This graph is called *resonance curve* of the oscillator.

The phenomenon of resonance can be seen easily: the oscillator absorbs most of the energy when the exciter frequency is equal to its natural frequency f_0. For both $f = 0$ as well as for $f \to \infty$, the energy current from the exciter to the oscillator becomes zero.

3. Even if not intended, the oscillator is damped by friction. If the damping effect is increased, for example by installing a damper, the resonance curve will change. The stronger the damping effect, the wider and flatter the hump of the graph, Fig. 2.5 and 2.6. Conversely, the curve becomes narrower the weaker the damping of the oscillator. If there is not damping at all, the resonance hump will become a very fine jag situated at the natural frequency, whose edges approach asymptotically an infinite value of the energy current.

> The energy current that flows from the exciter to the oscillator reaches its maximum value when the frequency of the excitation is equal to the natural frequency. Then, the oscillator is in resonance with the exciter. The stronger the damping effect, the flatter the resonance curve.

Exercise

1. Fig. 2.6 shows a section of Fig. 2.5. Compare the energy current values of the resonance curves for the frequencies $f = 1.5$ Hz, $f = 1.7$ Hz, $f = 2$ Hz.

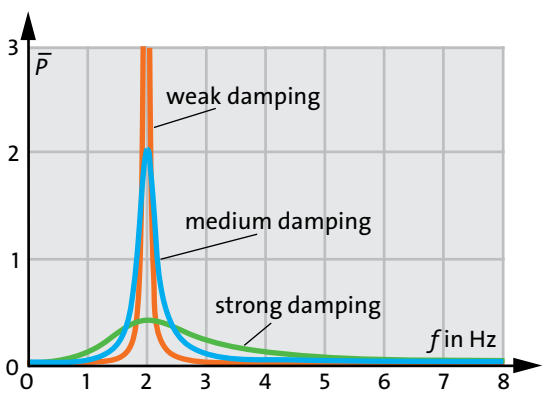

Fig. 2.5 Resonance curves for different dampings

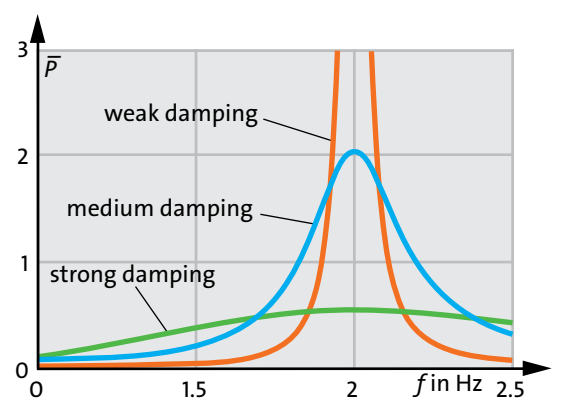

Fig. 2.6 Section of Fig. 2.5

2.3 How to draw a resonance curve

To draw a resonance curve, we have to measure the energy absorbed per time for many different frequency values. We do not have a meter for direct measurement of this energy. Therefore, we have to find some alternative way.

We try to take advantage of the relationship

$$P = \Delta v \cdot F_D$$

that we know from earlier observations. Here, F_D is the momentum current that flows through the damper and Δv is the difference of the velocities between the „connections" of the damper. As the left connection (a) does not move in Fig. 2.3, we can replace Δv by the velocity of the right connection (b). This velocity, however, is also equal to the velocity v of the oscillating body. We thus have:

$$P = v \cdot F_D \,.$$

We can set the momentum current F_D through the damper as proportional to the velocity:

$$F_D = k \cdot v \,. \qquad (2.1)$$

The energy current to the damper therefore becomes:

$$P = k \cdot v^2 \,.$$

As the velocity changes continuously in the course of a period, also this energy current changes periodically. With

$$v(t) = \hat{v} \cdot \sin(2\pi f t)$$

we obtain

$$P(t) = k \cdot \hat{v}^2 \cdot \sin^2(2\pi f t) \,.$$

We are now interested in the time average value of the energy current. We therefore have to find the time average of the term on the right side of the equation. k and $\hat{v}$ do not depend on time. Hence, we need to calculate the following

$$\overline{P}(t) = k \cdot \hat{v}^2 \cdot \overline{\sin^2(2\pi f t)} \,.$$

The time average of $\sin^2(2\pi f t)$ can be read directly from the function graph of the function

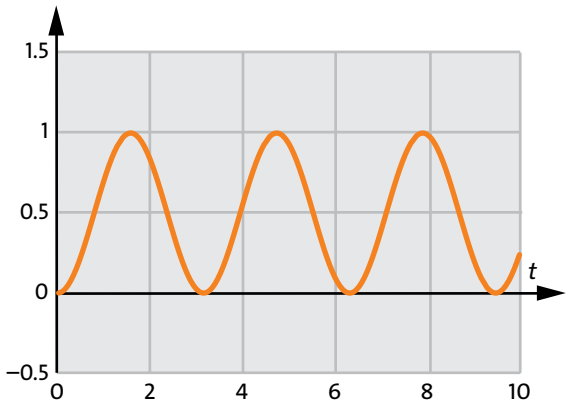

Fig. 2.7 Function graph of $f(t) = \sin^2(2\pi f t)$. The time average of the function values is 1/2.

$$f(t) = \sin^2(2\pi f t) \,,$$

Fig. 2.7. The curve is a sine curve that oscillates up and down between the values 0 and 1. The average value is obviously 0.5.

We can obtain the same result by means of the generally valid equation

$$\sin^2\alpha = \frac{1}{2} - \frac{1}{2}\cos(2\alpha).$$

that we have already used before. The mean value of the cosine on the right side is zero. Hence, only the addend 1/2 will be left.

We therefore have:

$$\overline{P} = \frac{k}{2} \cdot \hat{v}^2. \qquad (2.2)$$

The velocity amplitude and the damping constant k can both be measured easily. Hence, also the absorbed energy can be determined.

As $\hat{v}$ depends on the exciter frequency, also the average energy current depends on f. Thus, to draw the resonance curve, the velocity amplitude is measured for different values f and the absorbed energy is calculated on by means of equation (2.2).

Exercise

1. Earlier in this text, we stated that the average value of $f(t) = \sin^2(2\pi f t)$ is 0.5 and that this conclusion can be drawn from the function graph in Fig. 2.7. Explain this conclusion.

2.4 Resonance of a resonant circuit

With regard to experiments, an electric oscillator, i.e. a resonant circuit, has advantages and disadvantages. Although the measurements are easier and more precise – as no velocity and force sensors are required –, the oscillations can only be seen indirectly through the scales of a voltmeter, an ammeter or on the computer screen.

Fig. 2.8 shows a resonant circuit with a damping device and an „exciter". The exciter is an electric energy source that supplies an alternating voltage with a constant amplitude and whose frequency can be changed.

Just as in the case of the mechanical oscillator, „resonance curve" shall be understood as the function graph that illustrates the average value of the energy absorbed per time as a function of the exciter frequency f.

The resonance curve cannot be distinguished from that of a mechanical oscillator, Fig. 2.4 to Fig. 2.6.

Also here, the energy current can be determined by means of quantities that are easier to measure.

We apply

$$P = U \cdot I$$

to the resistor and replace the electric current that flows through the resistor by means of

$$I = \frac{U}{R}. \tag{2.3}$$

We obtain:

$$P = \frac{U^2}{R}.$$

As the voltage between the connections of the resistor changes according to

$$U(t) = \hat{U} \cdot \sin(2\pi f t)$$

we obtain

$$P(t) = \frac{\hat{U}^2}{2R} \cdot \sin^2(2\pi f t)\,.$$

From this the mean value of $P(t)$ can be derived in the same way as shown in the previous section:

$$\overline{P} = \frac{\hat{U}^2}{2R}. \tag{2.4}$$

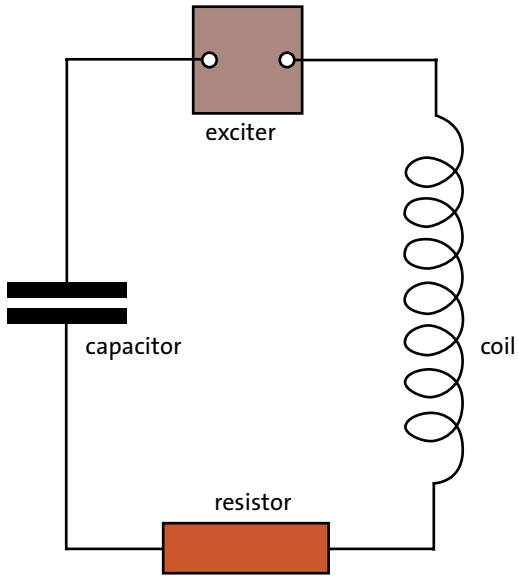

Fig. 2.8 Resonant circuit with exciter and damping resistor. The exciter creates a sine voltage with a constant amplitude. The frequency can be changed.

$\hat{U}$ and R can be measured easily. Therefore, also the absorbed energy can be determined without difficulty.

The equations (2.2) and (2.4) should actually be equivalent to each other. You might be confused that the constant k that characterizes the damper stands in the numerator whereas the constant R that characterizes the electric resistor stands in the denominator. The reason for this deviation is that k and R are not exactly equivalent to each other in the equations (2.1) and (2.3). While k tells us how well a mechanical resistor (the damper) *conducts* the momentum current, R is a measure for how strongly an electric resistor *hampers* the electric current. In fact, k corresponds to the reciprocal value of the electric resistance, i.e. a quantity that is also used in electrical engineering. It is called electric conductance.

2.5 Feedback oscillators

Systems that perform oscillations are technically important. They are used in clocks of any type. Also, they are needed for the creation of periodic waves that will be discussed in the next chapter.

In any case, we need a system that creates oscillations. The creation of oscillations is our problem now. „But haven't we just solved this problem?"" you might ask.

Let's recall once again how we defined „oscillations" at the very beginning: an oscillation: „…will run by itself after an initial energy supply" was stated there. But later we saw that all oscillations are damped. Mechanical oscillations lose energy due to friction. In case of electric oscillations, electric resistance causes the oscillation to gradually lose the energy that we put into it at the beginning. Every oscillation that we initiate therefore decays more or less quickly. To maintain an oscillation, the energy losses have to be compensated continuously.

We had already done that by means of an „exciter". Regarding our current problem, however, it is not a suitable method as the exciter itself has to contain an oscillation generator that can create a sineshaped momentum current or voltage. Hence, the problem would only be deferred.

We therefore have to manage to supply an oscillator with energy from a „normal" energy source, i.e. a source that does not already create an oscillation by itself.

We look at the oscillation from Fig. 1.9 as a specific example. The oscillator is illustrated once again in Fig. 2.9.

How could we supply energy to it? Let's try by pulling it to the right. The energy supply works – but only if we pull in the right moment, i.e. while the body is moving to the right. Due to the fact that we are pulling, the body gets additional momentum, the movement becomes more intense. But if we pull while it is moving to the left, we will slow it down (the absolute value of its momentum decreases) and we will take energy away from it. Hence, we are doing just the opposite of what we wanted. We see: to feed energy to the oscillation, we have to pull in the right moment, or rather in the right time interval.

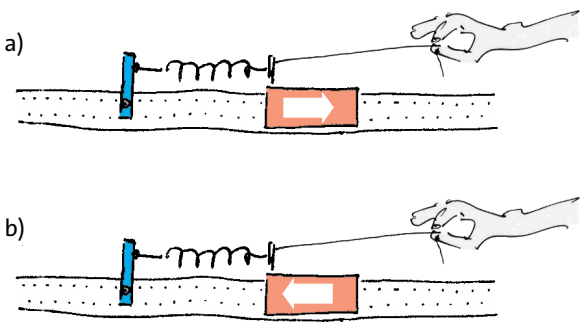

Fig. 2.9 If we pull on the thread while the body is moving to the right (a), we will provide energy for the oscillator. If we pull while it is moving to the left (b), we will take energy out of the system.

You will certainly find this observation logical if you think of how we can maintain – or also reinforce – the movements of a swing. The system has to be kicked at the right time.

If an oscillation is not maintained by a person who always pays attention to the pushing or pulling somewhere, we will need a device that ensures this process to occur automatically. Such devices exist, both for mechanical as well as for electric oscillations. Oscillations that are generated this way are called feedback oscillations. The oscillation itself controls the energy supply. We will not explain in detail how this is done because there are many different possibilities.

We keep in mind:

To maintain an oscillation, we need
- an oscillator
- an energy source
- a system to control the energy supply.

Exercises

1. How does the control system of the energy supply work in a pendulum clock? To make a resonant circuit perform a continuous oscillation, a „feedback system" is installed. How does that work? Use literature (for example an encyclopedia) or the Internet to find the information.
2. Also for periodic processes, which are not necessarily referred to as oscillations, a self-control is used: in the steam engine, the car engine, in a seesaw for children and also in the case of the back-and-forth movement of a tennis ball. Describe how the control system works in these cases.

3 SPECTRA

3.1 Some mathematical results

We will see that there are systems that can perform two or even more sine oscillations at the same time. To understand what this means and to see how such oscillations can be described, we have to familiarize ourselves with some mathematical results.

We have to add up sine functions, i.e. terms of the type

$$\hat{y} \cdot \sin(2\pi f t + \varphi)$$

Hence, the question is: what does the function

$$y(t) = \hat{y}_1 \cdot \sin(2\pi f_1 t + \varphi_1) + \hat{y}_2 \cdot \sin(2\pi f_2 t + \varphi_2) + \hat{y}_3 \cdot \sin(2\pi f_3 t + \varphi_3) + \ldots$$

look like? $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$, … are the amplitudes of the individual sine functions, f_1, f_2, f_3,... are their frequencies and φ_1, φ_2, φ_3,... the starting phases.

At first, it appears as if there was not much to say about this sum. A sine function is a simple, regular function. But if several of such functions are added up, the result will probably be something complicated and confusing.

But in fact, mathematics tells us that a few simple rules apply for such a sum; they will be addressed in the following.

We will not deal with the proof of these rules but trust that the conclusions from the mathematics books are correct. (If you would like to look things up in a mathematics book, you have to search with the keyword „Fourier series".)

The first rule we examine applies for a simple special case.

1. Sine functions with equal frequencies

At first, we only add up sine functions with the same frequency f. Amplitudes and starting phases can have any value. Hence:

$$y(t) = \hat{y}_1 \cdot \sin(2\pi f t + \varphi_1) + \hat{y}_2 \cdot \sin(2\pi f t + \varphi_2) + \hat{y}_3 \cdot \sin(2\pi f t + \varphi_3) + \ldots$$

Now, mathematics teaches us: the sum is again a sine function of frequency f. Fig. 3.1 shows an example. The three functions that are displayed in the image section (a) result in the function from image section (b) when added up.

> The sum of sine functions with equal frequencies f but any different amplitudes and starting phases is again a sine function with the frequency f.

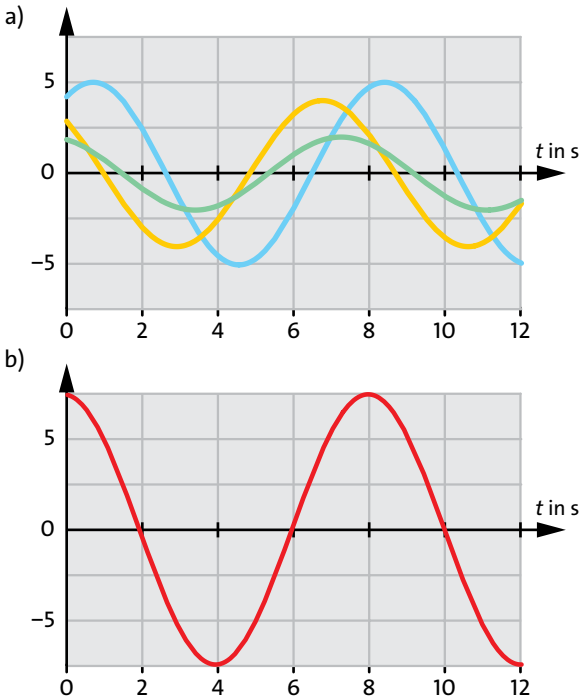

Fig. 3.1 Three sine functions with equal frequencies but different amplitudes and starting phases (a) result in a sine function with the same frequency (b) when added up.

2. Periodic functions

We add up two sine functions whose frequencies are integer multiples of the same *fundamental frequency* f_0. The frequency of the first sine function should be $n_1 \cdot f_0$, that of the second one $n_2 \cdot f_0$, where n_1 and n_2 are integers:

$$y(t) = \hat{y}_1 \cdot \sin(2\pi\, n_1 f_0\, t + \varphi_1) + \hat{y}_2 \cdot \sin(2\pi\, n_2 f_0\, t + \varphi_2)$$

During the period $T_0 = 1/f_0$ of the *fundamental oscillation*, the first addend performs n_1 oscillations and the second one n_2 oscillations. After the end of this period, each of the two addends has consequently performed a certain number of complete oscillations. Therefore, the sum function $y(t)$ is once again in the same state as at the beginning. In other words: the sum function is periodic with the period T_0.

Fig. 3.2 shows an example.

Here, the following addends were chosen:

$$5 \cdot \sin(2\pi \cdot 8\ \text{Hz} \cdot t + 0.5\,\pi),$$
$$3 \cdot \sin(2\pi \cdot 20\ \text{Hz} \cdot t + 1.5\,\pi).$$

The amplitudes are 5 and 3, the frequencies 8 Hz and 20 Hz and the starting phases are $0.5\,\pi$ and $1.5\,\pi$. The largest common divisor of 8 and 20 is 4. Hence, the frequencies are integer multiples of

$$f_0 = 4\ \text{Hz}.$$

We can therefore also write the addends as follows:

$$5 \cdot \sin(2\pi \cdot 2 \cdot 4\ \text{Hz} \cdot t + 0.5\pi),$$
$$3 \cdot \sin(2\pi \cdot 5 \cdot 4\ \text{Hz} \cdot t + 1.5\pi).$$

The period of the sum function is

$$T_0 = 1/f_0 = 0.25\ \text{s}.$$

During this time, the first addend performs 2 oscillations, the second one 5 oscillations, see image section (a). The image of the sum function is a curve section that repeats itself every 0.25 seconds; see image section (b).

You certainly anticipate that the rule we have just discovered can be generalized. Even if we add up more than two sine functions whose frequencies are integer multiples of a frequency f_0, the result will be a periodic function.

Not as obvious, on the other hand, is the following generalization that goes even further:

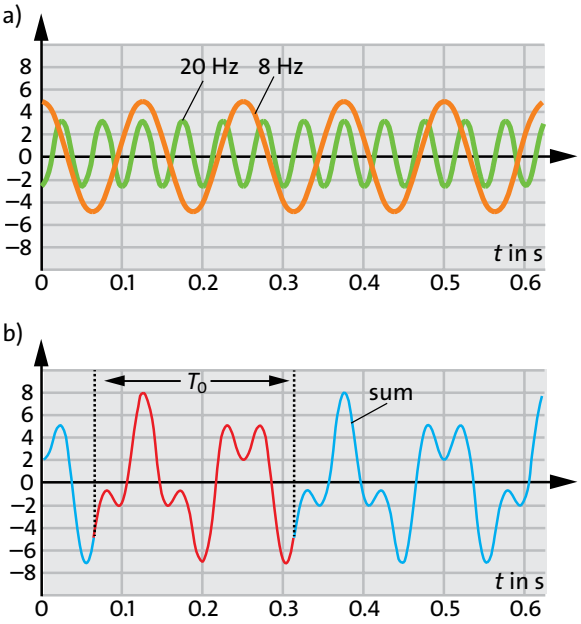

Fig. 3.2 Two sine functions whose frequencies are integer multiples of the same fundamental frequency f_0 (a) result in a periodic function with the period $T_0 = 1/f_0$ (b) when added up. In the image section (b), the curve section that repeats itself periodically is highlighted in a red color.

Every periodic function (frequency $f_0 = 1/T_0$) can be expressed as a sum of sine functions whose frequencies are integer multiples of f_0.*

Once again the meaning of this phrase in other words: there is a periodic function $y(t)$ with the period $T_0 = 1/f_0$. This function can be written as a sum of sine functions:

$$\begin{aligned} y(t) = {} & \hat{y}_1 \cdot \sin(2\pi \cdot 1 \cdot f_0 \cdot t + \varphi_1) \\ & + \hat{y}_2 \cdot \sin(2\pi \cdot 2 \cdot f_0 \cdot t + \varphi_2) \\ & + \hat{y}_3 \cdot \sin(2\pi \cdot 3 \cdot f_0 \cdot t + \varphi_3) \\ & + \hat{y}_4 \cdot \sin(2\pi \cdot 4 \cdot f_0 \cdot t + \varphi_4) \\ & + \ldots \end{aligned}$$

Of course, it can be that very large number of addends is needed. But if we content ourselves with representing the function approximately, a few addends will be sufficient.

3. Any function

The rule that we have just learned is already astonishing enough. But its validity is actually not limited to periodic functions. Also nonperiodic functions can be written as a sum of sine functions. However, the frequencies of the addends will then no longer be integer multiples of a fundamental frequency. The following rule applies:

Every function can be expressed as a sum of sine functions.** We summarize the results * and **:

> Every function $y(t)$ can be expressed as a sum of sine functions.
>
> If $y(t)$ is periodic (period T_0) the frequencies of the sine functions are integer multiples of the fundamental frequency $f_0 = 1/T_0$

This is an important result because we often want to know which sine components a function is made of.

Now, this result is not very useful as long as we do not know how its sine components can be determined. Let's assume that there is a given time function. What are its addends? What are the values of the amplitudes and of the phases of the different addends?

Of course, mathematicians are able to calculate these amplitudes and phases, but the method is quite complicated. We therefore use a method that is less sophisticated but more practical instead: we make a computer solve the problem. This means that we have the work done by a program that uses the mentioned mathematical method. In doing so, we also facilitate our task in another way: we would only like to know the amplitudes but not the phases of the sine functions. Hence, we ask: „What is the contribution of the sine function with a given frequency in a given time function?". We do not ask about the starting phase, meaning that we do not inquire about the position of the sine function on the t-axis.

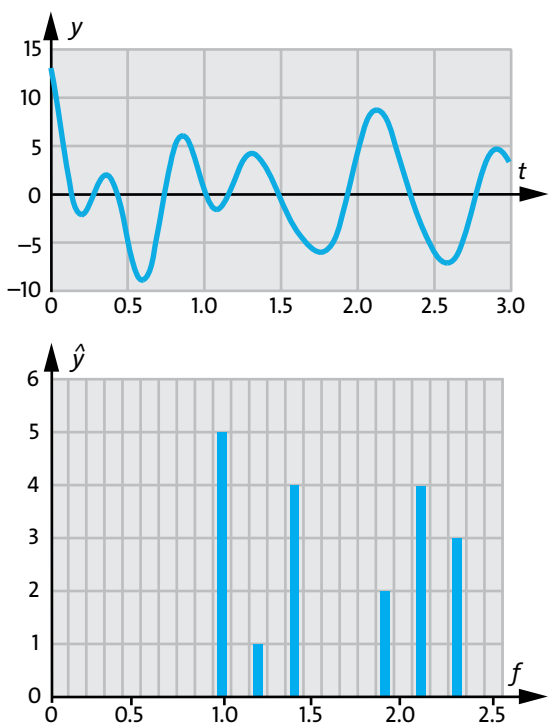

Fig. 3.3 Top: time function; bottom: y associated spectrum

Exercises

1. Calculate (with the calculator or the computer) for $0 < x < 4\pi$ a table with the values of the following function: $y(x) = \sin(2x) + 3\sin(2x + \pi/2) + 3\sin(2x + \pi)$. Represent the function graphically.
2. Calculate (with the calculator or the computer) for $0 < x < 4\pi$ tables with the values of the following functions:

$$y(x) = \sin x,$$

$$y(x) = \sin x - \frac{1}{9}\sin 3x,$$

$$y(x) = \sin x - \frac{1}{9}\sin 3x + \frac{1}{25}\sin 5x,$$

$$y(x) = \sin x - \frac{1}{9}\sin 3x + \frac{1}{25}\sin 5x - \frac{1}{49}\sin 7x.$$

   Represent the function graphically. Could you make a conjecture?
3. Calculate (with the calculator or the computer) for $0 < x < 4\pi$ a table with the values of the following function:

$$y(x) = \sin x + \frac{1}{3}\sin 3x + \frac{1}{5}\sin 5x + \frac{1}{7}\sin 7x + \frac{1}{9}\sin 9x + \frac{1}{11}\sin 11x.$$

   Represent the function graphically. Could you make a conjecture?

3.2 Spectra

We start with any time function whose sine components are known. We would like to tell someone else about this known fact. How can that be done? A somehow bold method would be to draw a table of values with the frequencies of the existing sine terms in the first column and the corresponding amplitudes in the second column. If we were interested in the starting phases, we could list them in a third column. But such a table would not be very clear. A graphical display is more suggestive.

In fact, it is a common practice to illustrate the sine components of a function $y(t)$ in a graph: the amplitude of the sine functions over the frequency.

We are therefore dealing with two different graphical illustrations:

- with the time function $y(t)$,
- with the spectral function $\hat{y}(f)$.

The graphical display of the spectral function is called spectrum of the function.

With some mathematical skills and understanding of physics, you will be able to draw the spectrum for a given function $y(t)$. You will learn it best from the applications that are discussed in the following sections. For the time being, we will only show one single example of a time function together with its spectrum: Fig. 3.3. As $y(t)$ only consists of 6 sine functions with frequencies whose values are far apart from each other, it is useful to draw the spectrum as a bar chart. If the spectral function shows a steady progression, it will be drawn like a normal function graph.

Exercise

1. Graphically illustrate the spectra of the functions from Exercises 2 and 3 of the previous section. Use $x = 2\pi t$.

3.3 Double oscillators

Fig. 3.4 shows once again an oscillator that we know from earlier chapters. We assume that the oscillating body has a mass of 0.5 kg and the spring a spring constant of 30 N/m.

We make the body oscillate again, but record its position with a sensor this time. The digital signal provided by the sensor is transferred to the computer, which will then determine the sine contribution to the movement by means of an appropriate program. It will display the result on the screen in form of two graphs:

1. The position of the body as a function of time: $s(t)$
2. The spectrum of this function: the abscissa is the frequency, the ordinate is the amplitude $\hat{s}$ of the sine components of the time function $s(t)$.

The task is still so simple that we already know the result in advance. We can therefore check whether the computer works correctly. It does indeed work correctly. The result it provides is shown in Fig. 3.5. The time function of the movement of the body is a sine function. The spectrum shows a pointed jag, a „peak".

After having learned how to deal with the computer and how to interpret the graph, we will now address more interesting movements. We set up a double oscillator: two bodies and three springs as illustrated in Fig. 3.6.

Playing a bit with this oscillator gives us the impression that it is substantially different from that in

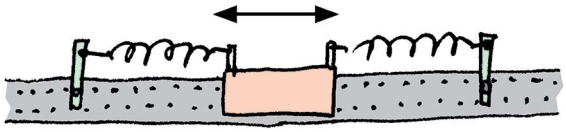

Fig. 3.4 The spectrum of the function ***x*(*t*)** (position of the glider as a function of time) only contains a single sine contribution.

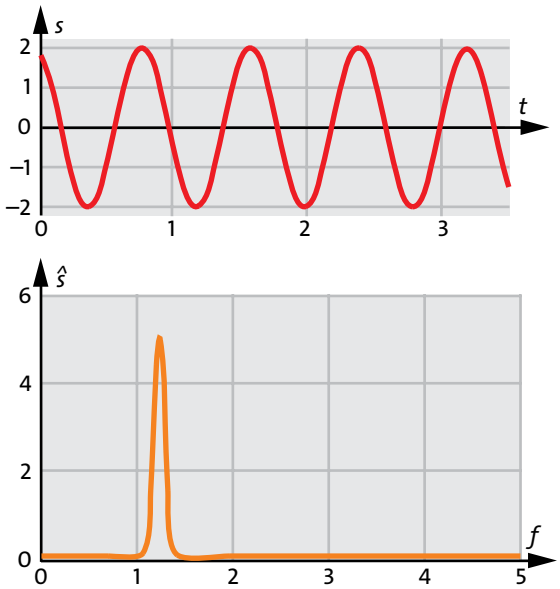

Fig. 3.5 Time function (top) and spectral function (bottom) for the movement of the glider from Fig. 3.4

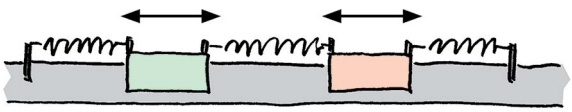

Fig. 3.6 Double oscillator. If only one of the two natural oscillations is excited, both bodies will perform a pure sine movement with equal frequencies. If both natural oscillations are excited at the same time, the position of each of the two bodies will be described by the sum of two sine functions. For each of the two bodies of a double oscillator, the function $s(t)$ is the sum of two sine functions.

Fig. 3.4. The oscillator from Fig. 3.4 always performs the same movement, no matter how it is nudged. Of course, it can oscillate with a smaller or a larger amplitude, but it will always perform a sine-shaped movement. Our double oscillator seems to behave differently. Depending on how it is nudged, it makes a different movement. Also, we can see that in general it does not perform any sine movement.

Now we will examine the double oscillator a bit more systematically by recording the time function

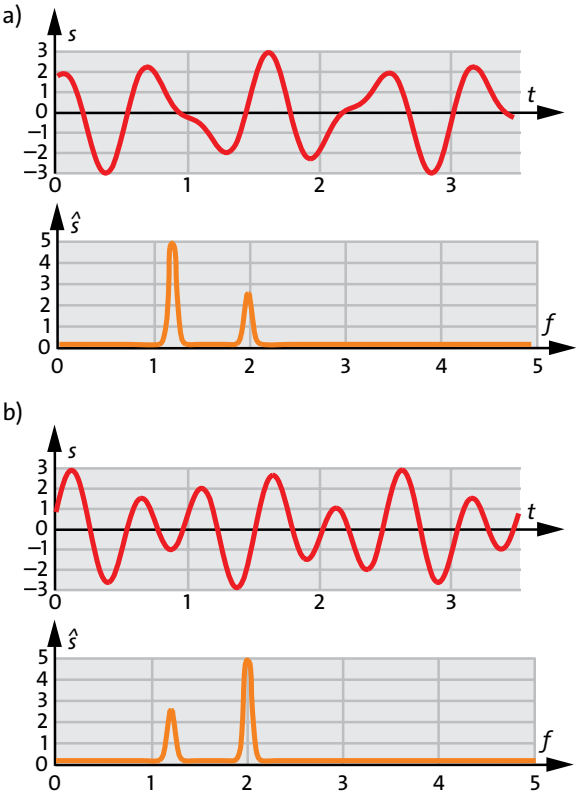

Fig. 3.7 The position of the body on the left as a function of time, recorded twice. The oscillator was nudged differently in the two cases. The associated spectrum is shown respectively below.

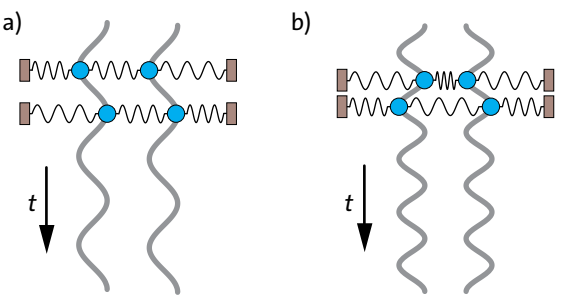

Fig. 3.8 Position of the two bodies as a function of time, (a) for the first natural oscillation, (b) for the second natural oscillation

and the spectrum. We mount the sensor on one of the two gliders and nudge the bodies in any way so that they start moving. The computer displays the result. It could look like that shown in Fig. 3.7a.

We repeat the experiment. Again, we nudge the bodies but this time a bit differently so that we obtain a different result, for example the one from Fig. 3.7b. We repeat the experiment a few more times and observe the result displayed on the screen each time. If we concentrate exclusively on the time functions, our first impression will be confirmed: depending on how the bodies are nudged, we obtain a different function curve. We cannot identify any systematic scheme. But we will get a completely different impression if we look at the spectra. In any case, the spectrum consists of two peaks that are always at the same frequencies f_1 and f_2. In the example of our Figures, we have the value $f_1 = 1.2$ Hz and $f_2 = 2.0$ Hz. We can conclude from this observation that the time function is the sum of two sine functions in any case. The various experiments differ from each other in the amplitudes of the two sine functions that have different values each time.

Now we repeat the series of experiments but put the sensor on the other body. We try to nudge the bodies in the same way as in the first series of experiments. We find that the time function that describes the movement of the second body is also a sum of two sine functions. Again, the frequencies are f_1 and f_2, i.e. the same as in the first series of experiments. The amplitudes, in turn, have other values than before.

> For each of the two bodies of a double oscillator, the function $s(t)$ is the sum of two sine functions.

The frequencies f_1 and f_2 of the two sine functions are called natural frequencies of the double oscillator.

We have seen that the amplitudes of the two sine functions depend on how the oscillator is nudged. Now we would like to examine this influence.

We therefore nudge the double oscillator in a very specific way: the two bodies are displaced leftwards by the same distance and then released simultaneously. You can probably predict what happens next: both bodies oscillate in a sine-shaped way with the frequency f_1, i.e. they oscillate in sync and the distance between them will always remain the same. Fig. 3.8a shows the positions of the two bodies as a function of time. This result does not conflict with our experiences up to present. Only the amplitude of the oscillation with the frequency f_2 is zero.

Now we nudge the double oscillator in a different way: both bodies are displaced outwards by the same distance and then released simultaneously. Again, both perform a pure sine movement but this time against each other. The frequency is f_2, Fig. 3.8b.

The two sine movements that we have excited are called *natural oscillations* of the system.

A double oscillator can perform two different natural oscillations. A natural oscillation is characterized by the fact that each of the two bodies makes a pure sine movement.

If we nudge the double oscillator in any random way, just as we did at the beginning, so that the time functions $s(t)$ of the two bodies are composed of two sine functions, we can say that the system performs two natural oscillations at the same time.

Exercises

1. We have observed that the first of the two natural oscillations that we had triggered has a lower frequency than the second one. This could also have been predicted.
2. We are interested in the result of an experiment that is similar to that from Fig. 3.6. However the springs are no longer identical. (a) The two external springs are quite hard, that in the middle is very soft. What can be said about the two natural frequencies? (b) The spring in the middle is hard, the two external springs are soft. What can be said about the two natural frequencies?
3. In the first oscillator that we have examined, Fig. 1.3, two bodies are moving as well. However, we did not notice anything about a second frequency back then. Can you comment on this matter?
4. How can two pendulums that are suspended next to each other be transformed into a double oscillator?

3.4 Multiple oscillators

Now we connect not only two, but three, four or even more bodies to each other by means of springs, Fig. 3.9. The experiment confirms what we could also have guessed. The spectrum of the triple oscillator has three peaks, i.e. also three natural frequencies whereas the spectrum of a quadruple oscillator has four peaks, i.e. four natural frequencies, etc.

The spectrum of a N-fold oscillator has N peaks. The system has N natural frequencies and can perform N different natural oscillations.

Again, we recognize a natural oscillation by the fact that all the bodies perform pure sine movements.

The question of how the N bodies move in case of the different natural oscillations and how to excite the individual natural oscillations is more difficult to answer. There are rules and principles, but they are complicated and should not be of interest here. With some physics skills, however, we can guess easily how at least some of the natural oscillations look like.

Exercises

1. Describe the movement of the individual bodies for a maximum number of natural oscillations (a) for a triple oscillator, (b) for a quadruple oscillator.
2. Describe the movement of the individual bodies for one of the natural oscillations of a thousand-fold oscillator.

3.5 When inertia and elasticity are no longer separated

The oscillators that we have examined so far had a common feature that we have taken for granted: they consisted of „bodies“ and „springs“. We have imagined the bodies as similar to hard blocks. They have a certain mass and therefore they have inertia but they are not elastic or deformable. Regarding the springs, by contrast, we have assumed its mass to be so small that it could be neglected. We considered its elasticity as its only relevant property. Let's take another look at the oscillator from Fig. 1.2: the oscillating ruler that is clamped on one side. Here, inertia and elasticity are not separated from each other. Each piece of the ruler is both inert as well as elastic, and both characteristics are relevant. The ruler can practically not be distinguished from an oscillator that consists of a very large number of very small inert bodies that are separated from each other and that are connected to one another by means of many tiny springs. If things worked this way, we could in fact expect the ruler to have not only one natural oscillation – as it appeared at first – but several ones, or rather many of them. And this is actually true.

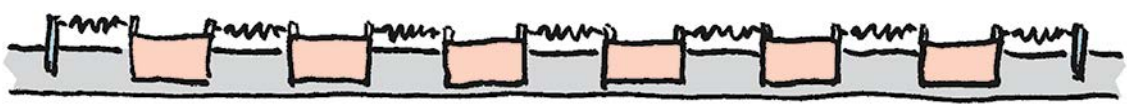

Fig. 3.9 The six-fold oscillator can perform six different natural oscillations.

Such natural oscillations can be observed particularly well in a slightly different system, which, however, behaves essentially like the ruler: an elastic rope that is clamped on both sides, Fig. 3.10.

To find out the frequencies of the natural oscillations, we move one of the two fixations sinusoidally up and down with a small amplitude. We start with a movement at a very low frequency and gradually increase the frequency. Exactly the same is done to draw the resonance curve of the system.

Now we find that the rope makes a very simple movement at a specific frequency: the shape of the rope changes between two extremes. Each point of the rope makes a pure sine movement; i.e. we are dealing with a natural oscillation. By the way: the amplitude of the movement is very large. Hence, the rope is in resonance with the exciter. If we increase the frequency further, the amplitude will decrease again and the movement will no longer be sine-shaped. By further increasing the frequency, we will arrive again at a state in which the individual points of the rope perform a sine movement with a large amplitude. The shape of the rope, however, is now different from what it was during the first resonance. We have a second natural oscillation and a second resonance. We can increase the frequency more and more and will observe a series of natural oscillations.

The natural oscillation with the lowest frequency is called *fundamental oscillation*; the other ones are referred to as *harmonics*. The frequencies of the harmonics are integer multiples of the fundamental frequency.

We can also try to excite the natural oscillations with our hands. But the result will be somewhat poor.

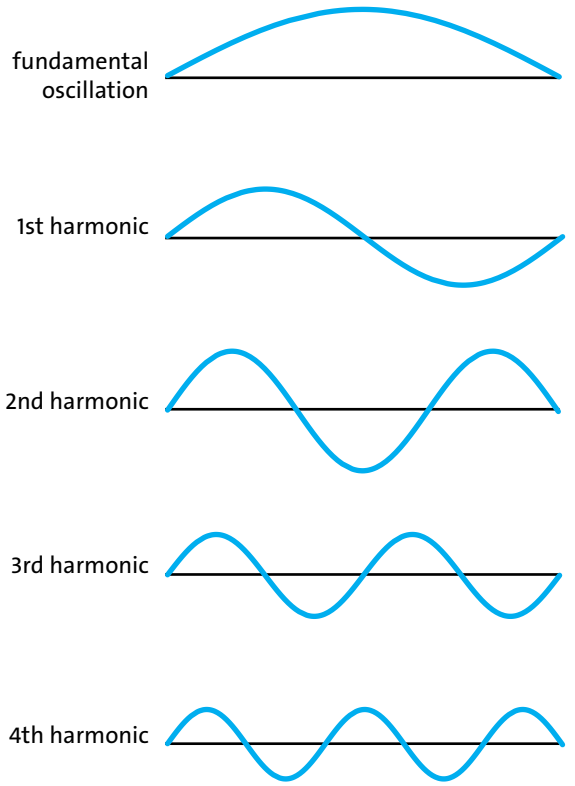

Fig. 3.10 Snapshots of the natural oscillations of a taut rope or the string of a musical instrument

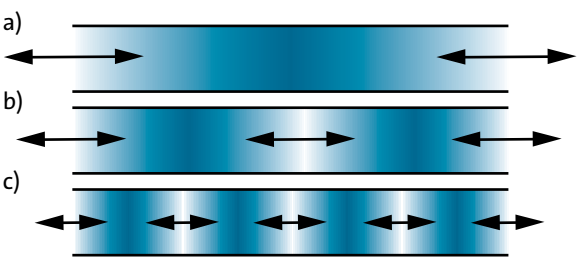

Fig. 3.11 Air oscillations in a wind instrument (flute, clarinet, organ pipe...). (a) fundamental oscillation, (b) first harmonic, (c) fourth harmonic. The arrows indicate how the air is moving. The density of the air changes due to this movement. The gray shading illustrates a sort of „snapshot" of the density.

Musical instruments

You might have noticed that the structure of our experiment is similar to a string instrument. To achieve higher frequencies, the oscillating „cord" in a musical instrument is a steel wire. When playing the instrument, the string is excited in different ways to perform oscillations: by means of beating or plucking with the fingers in case of guitars, through beating with a small felt hammer in the case of pianos, by means of striking with the bow in case of violins, violas, cellos and basses. Apart from the fundamental oscillation, several or many other harmonics are excited in any case.

The working principle of the wind instruments is not very different from that of the string instruments. But there, not a string but the air inside the instrument performs the oscillations. While every piece of a string oscillates back and forth transversally to the string direction, the air in a wind instrument moves in a longitudinal direction, Fig. 3.11. This class of instruments includes the recorder and the transverse flute, the clarinet and the saxophone, the oboe and the bassoon, the brass instruments trumpet, trombone, horn, etc., and also the organ.

The spectra of the sounds of wind instruments are very similar to the spectra of sounds of the string instruments although these sounds are created very differently. Also in this case, the frequencies of the har-

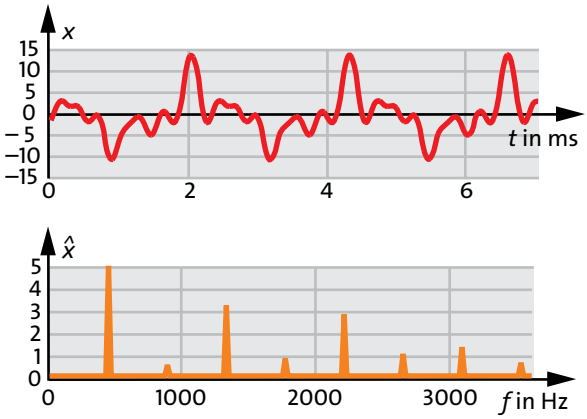

Fig. 3.12 Oscillation spectrum of a clarinet. The frequencies of the natural oscillations are integer multiples of a fundamental frequency (here the standard pitch A with $f = 440$ Hz).

monics are integer multiples of a fundamental frequency.

Fig. 3.12 shows the movement of the air in a clarinet at the top as well as the associated spectrum below. It is obvious that the frequencies of the harmonics are integer multiples of the fundamental frequency.

Also objects with other shapes can be excited to perform oscillations, and such objects have their characteristic natural oscillations. In most cases, however, the spectrum is much more irregular than in the case of a string or wind instrument: the frequencies of the harmonics are no longer integer multiples of the fundamental frequency. As the frequency of the fundamental oscillation is generally in the range in which our ear is sensitive, we can hear all these oscillations. They are the sounds that constantly surround us and that are formed when two objects collide.

Some musical instruments also have spectra with harmonic frequencies that are no integer multiples of the fundamental frequency. They are the ones for which the oscillating body is not only extended in one direction: the cymbal, the timbale and the bell.

The quartz clock

The name is due to the fact that, instead of the pendulum or the balance wheel, the clock has a tiny quartz crystal that performs one of its natural oscillations and therefore sets the clock pulse.

The fact that it has to be a quartz crystal is due to a special property of this material. When a quartz crystal is deformed – compressed or stretched – it charges itself electrically on two opposite faces. Conversely, a crystal changes its shape when it is charged on two opposite sides by applying a voltage.

Thanks to this property, the crystal can be provided with the energy that it loses while oscillating due to the inevitable damping effect. This way, an electric circuit can at the same time be given a time signal, which repeats itself with each period of the quartz, with the mechanical oscillation. This time signal is used to control the clock hands.

4 WAVES

Everyone knows them as water waves, Fig. 4.1: a sort of deformation of the water surface that moves by itself.

Another wave phenomenon is the *sound*: very small pressure changes of the air that move through the air.

Water and sound waves are so interesting that it would definitely be worth dealing with waves in greater detail.

In fact, however, waves have a much higher significance in science and technology because many phenomena of which we would not expect it in advance can be interpreted as waves.

At first and above all, there is the large class of the so-called electromagnetic waves. They include the waves that are used for radio and television broadcasting, for mobile phones and for wireless landline phones. In addition, there are the „microwaves" used in microwave stoves. Then, there is a wide range of phenomena that we call radiations: infrared and ultraviolet radiation, X-rays and gamma radiation, and finally the light that we are all familiar with. Later we will deal with the question about the medium in which these waves are moving. Comparing these waves to water waves, what would be the equivalent of the water and what would correspond to the deformation of the water?

Another type of waves, i.e. the *gravitational waves,* are much more difficult to create and to detect. In principle, they are formed when several masses move opposite to each other. But to be intense enough as to be detectable, the masses need to have huge values. Noticeable gravitational waves emerge for example during the explosion of a star; i.e. a process that is called supernova.

What's more, there is also another, maybe even more important type of waves in nature: the *matter waves*. All matter can be described as a wave. *Quantum mechanics* teaches us the meaning of this statement that seems almost absurd.

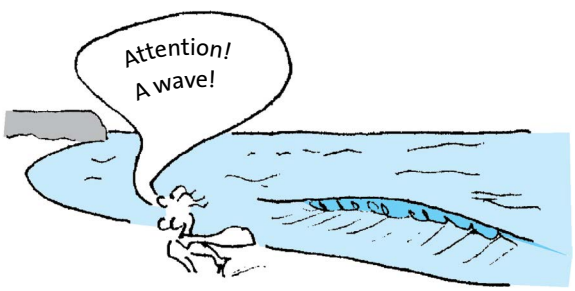

Fig. 4.1 Water wave

Waves transport energy. Having bathed in the sea during a heavy swell, you might have already felt yourself that water waves carry energy. During an explosion, windows can break even at a long distance. The energy is transmitted with sound waves in this case. Also the movements of the Earth during an earthquake are sound waves. You certainly know that light and other electromagnetic waves carry energy.

Now you will understand that it is worth dealing with waves. At first, we will omit the gravitational waves and the matter waves.

4.1 The carrier of waves

We start our examination with a very simple type of wave: „waves on a string".

A long string is laid on the floor and one of its ends is moved briefly and forcefully upwards and back downwards immediately after. A wave moves away

from the end that we move. Fig. 4.2 shows 3 snapshots.

Hence, what we call a wave is the deformation that runs through the string. It is logical that no such a wave can be created without a string.

We generalize: a condition for the existence of a wave is something in which the wave can move. We call this „something" *carrier of the wave*. (In our case, the string is the carrier.) The carrier is at first in its *ground state*. (The string is extended straightly on the floor.) A generator of the wave or *sender* then causes a change of the ground state for a short time. (A person moves the string up and down at one point during a short time.) This change of the ground state moves through the carrier or beyond the carrier.

With these considerations in mind, we examine a second wave type: a water wave in a long gutter. The water is the carrier of the wave. In the ground state, the surface of the water is horizontal everywhere, Fig. 4.3a. We dip a body briefly into the water on one end of the gutter, Fig. 4.3b, and pull it back out. This creates a deviation from the ground state: a bulge and a dent are formed on the surface of the water and move away from the point of creation, Fig. 4.3c.

A third example of a wave is shown in Fig. 4.4. The carrier of the wave is a long and slightly pre-stressed steel spring. The change of the ground state: one end of the spring is quickly moved a bit to the left and immediately back to its old place. Also this perturbation of the ground state moves away from the point of creation by itself.

We summarize:

> A wave needs a carrier. At the point of creation of the wave, the ground state of the carrier is changed quickly. This deviation from the ground state moves away by itself.

In all examples that we have examined, the wave carrier was moving besides the wave. Please distinguish these two movements properly!

The movement of the wave carrier is sometimes transversal to the direction of propagation of the wave, for example in the case of the wave on a string. Such waves are called *transverse waves*.

In other cases, the back-and-forth movement of the wave carrier goes in the same direction as the movement of the wave, for example in case of the wave in the steel spring. Such waves are called *longitudinal waves*.

But there are also waves for which the movement of the carrier is more complicated. In a water wave, for example, the carrier, i.e. the water, moves on a closed curve.

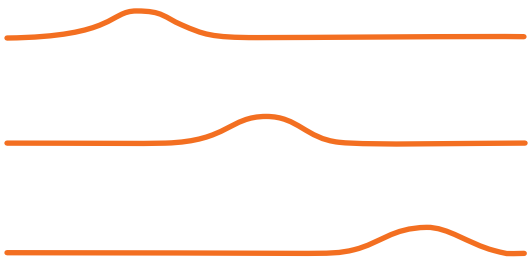

Fig. 4.2 Three snapshots of a wave on a string

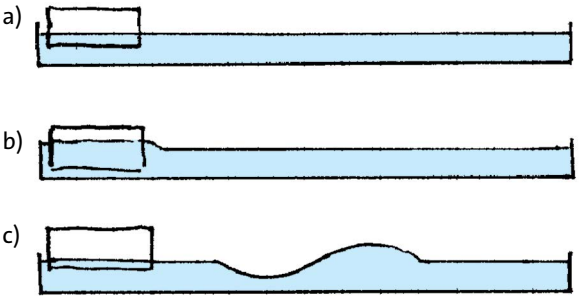

Fig. 4.3 The body that swims in the water (a) is pushed downwards for a short time and pulled back up (b). A wave moves away from the „sender" (c).

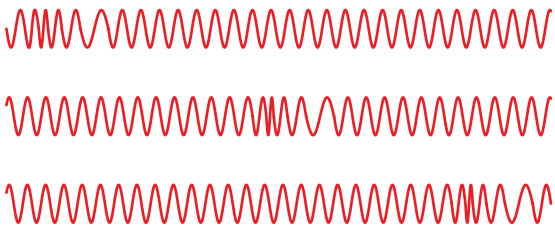

Fig. 4.4 Three snapshots of a longitudinal wave in a steel spring

And sometimes nothing moves at all, for example in the case of the electromagnetic waves.

> Do not confuse the movement of the wave and the movement of the wave carrier.

4.2 The velocity of waves

Try to make a wave on a string faster or slower by moving the string end in different ways. All you will achieve is to change the shape and the size of the deformation of the wave. You do not have any influence on

the velocity. It is different from throwing a stone. The velocity of the stone depends on the momentum that the stone has received in the process of throwing. But what does the velocity of a wave depend on?

The answer to this question is actually a bit intricate. But we can already give you an approximate answer at this point:

> The velocity of a wave depends on the carrier in which the wave is moving.

A wave on the surface of water has a different velocity than on the surface of alcohol, gas or mercury. Or a wave in a hard spring moves with a different velocity than in a soft one. Sound waves move in air at approximately 300 m/s, in water at 1480 m/s. (The waves in water can be created easily in a swimming pool by diving and letting out a yell under water.)

Light moves in the so-called empty space (we will see later that the empty space is actually not really empty) at 300 000 km/s but only at 200 000 km/s in glass.

Exercise

1. Many dominoes are put next to each other in an upward position. If we nudge the first one, it will fall against the second domino so that it will fall as well, etc. Hence, a change of the state of the dominoes runs through the row. What does this process have in common with a wave? How does it differ from a wave?

4.3 One-, two- and three-dimensional wave carriers

We would like to sort the waves that we have come across so far according to the number of dimensions of the wave carrier.

A wave on a string moves on a string, and a string can be regarded as a one-dimensional object. Water waves move on the two-dimensional surface of water. Sound waves are finally an example for a three-dimensional wave carrier.

If we want to track the progression of a one-dimensional wave, we will have to focus our view on one point, e.g. a maximum of deviation. In the case of a two-dimensional wave, the peaks are lines. The movement of the wave can therefore be observed by looking at such a line. These lines are called *wave fronts*. For three-dimensional waves, the wave fronts are surfaces.

Wave fronts can have a variety of shapes. However, we will often come across waves for which they have a simple shape.

The simplest two-dimensional waves are those whose wave fronts are straight lines, Fig. 4.5. For the simplest three-dimensional waves, the wave fronts are planes. Therefore, they are called *plane waves*.

Another simple wave type among the two-dimensional waves is the circular wave, Fig. 4.6. Such a shape is formed when the wave moves away from the sender in all directions. For example, they are created on a water surface by throwing gravel stones into the water. The equivalent for three-dimensional waves is a *spherical wave*. An example is the sound wave that moves away from a bursting balloon.

As the energy current of the wave spreads over a constantly growing spherical surface, the maximum deviation of the wave decreases with a growing distance from the sender. A small section of the spherical wave can be regarded approximately as a plane wave.

Fig. 4.5 Two-dimensional wave with a straight wave front

Fig. 4.6 Two-dimensional wave with a circular wave front

4.4 Sine waves

We are not satisfied with a single wave moves over the water. We would like the transport process to continue. What can we do? Well, we can simply send out several or many individual waves, i.e. a *wave train*. To create a wave train, it is not sufficient that the body, which generates the waves, is moved up and down once. We have to move it up and down time and again, i.e. periodically.

In each of our examples, the sender had to perform a specific type of movement: an up-and-down movement for a water wave and a back-and-forth movement for the wave in the steel spring. Hence, to create a periodic wave, this movement has to be periodic.

If this periodic movement is sine-shaped, a *sine wave* will be result.

A sine wave in a string can be detected easily as a snapshot of the string has precisely the shape of a sine function graph. We imagine making several snapshots of the string shortly after each other, Fig. 4.7. The first picture shows a sine shape, the second one as well, the third one too, etc. But the three pictures are offset against one another.

In mathematical terms, this behavior of a wave on a string is described by the function

$$y(x,t) = \hat{y} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right]. \qquad (4.1)$$

Here, x is the position coordinate in the direction of propagation of the wave. t is the time. The dependent variable y will later stand for a variety of physical quantities: for the water level above the normal level in the case of a water wave, for the pressure or the velocity of the air in case of a sound wave, for the electric or the magnetic field strength in case of an electromagnetic wave. For the wave on a string that we are currently talking about, y is the *deviation*: the displacement of the string from the resting position, transversally to the direction of propagation of the wave. In case of a longitudinal wave as in Fig. 4.4, the deviation y has the same direction as the direction of propagation.

Just as in the case of an oscillation, $\hat{y}$ is also called *amplitude* here.

The special feature of the function (4.1) is the fact that y depends on *two* independent variables: on the position x and on the time t. Making a snapshot of the wave at a specific instant of time t_1 means inserting a specific value t_1 in equation (4.1) for the time. The result is a function that only depends on a single independent variable, i.e. on x:

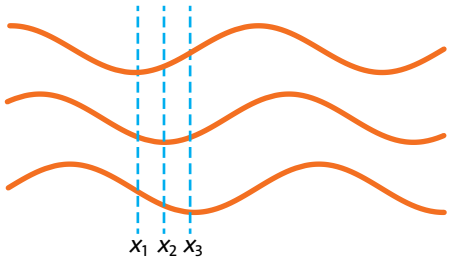

Fig. 4.7 Three snapshots of a rope through which a sine wave is moving

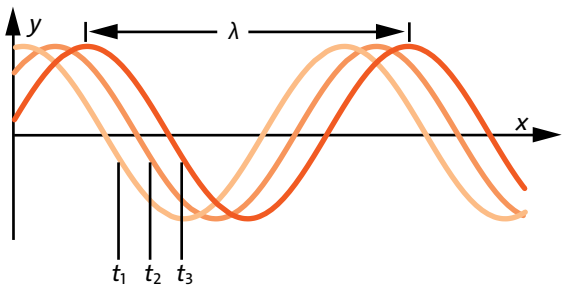

Fig. 4.8 The functions that correspond to the three instants of time t_1, t_2 and t_3 only differ from each other in the starting phase.

$$y_{t1}(x) = \hat{y} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t_1}{T}\right)\right].$$

The functions that correspond to snapshots at the later times t_2 and t_3 are

$$y_{t2}(x) = \hat{y} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t_2}{T}\right)\right],$$

$$y_{t3}(x) = \hat{y} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t_3}{T}\right)\right].$$

Fig. 4.8 shows the graphs of the three functions $y_{t1}(x)$, $y_{t2}(x)$ and $y_{t3}(x)$. We can see that the choice of the instant of the snapshot only has an influence on the starting phase of the function. For the picture, this means that the sine line moves to the right over time.

„Snapshot“ means: choose an instant of time, for example t_1, and consider y as a function of the position, i.e. the function $y_{t1}(x)$. We would now like to reverse things. We choose a fixed position x_1 and let the time run:

$$y_{x1}(t) = \hat{y} \cdot \sin\left[2\pi\left(\frac{x_1}{\lambda} - \frac{t}{T}\right)\right].$$

This is the equation of a sine oscillation. Its starting phase is $2\pi \cdot x_1/\lambda$. In other words: at each position x, the

wave carrier performs a sine movement, for example at the point x_1 or at the point x_2 or x_3 in Fig. 4.7. The function graphs of the corresponding movements are displayed in Fig. 4.9.

The distance between two adjacent peaks of the graph in Fig. 4.8 corresponds to the *wavelength* λ. The distance between two peaks of the graph in Fig. 4.9 corresponds to the period T.

You have seen that it is quite time-consuming to describe the function (4.1). If you have more trust in the computer than in mathematics, you can convince yourself much more easily that equation (4.1) describes a wave. Just enter the function in an algebra program and have it displayed as an animation. If only the qualitative progression is relevant, it will be sufficient to enter the function

$$y = \sin(x - t) .$$

(Your computer possibly requires other names for the variables.)

Equation (4.1) only describes a one-dimensional wave, for example a wave on a string. But there are also two- and three-dimensional sine waves. The position coordinate x will then be measured in the direction of propagation of the wave. This means that the wave fronts have to be lines for a two-dimensional wave and planes for a three-dimensional wave. Fig. 4.10 shows a wave on a two-dimensional wave carrier.

A circular wave (on a two-dimensional carrier) or a spherical wave (in a three-dimensional carrier) can never be real sine waves. Even if the sender makes a sine movement, the maximum deviation, i.e. the „amplitude“, will decrease towards the outside, Fig. 4.11.

Exercises

1. We look at the sine wave on a water surface. The x-coordinate in the wave equation (4.1) has to be measured in the direction of propagation of the wave. Now we turn the coordinate system in a way that the x-direction is parallel to the wave fronts. How will equation (4.1) change?
2. Water waves on the sea or on a pond are sometimes nearly sine shaped. Which wavelengths can be found?
3. Compare the „snapshots“ of the function of the equation (4.1) for $t = 0 \cdot T$, $t = 1 \cdot T$, $t = 2 \cdot T$ and $t = 3 \cdot T$. Explain the result.
4. Also the equation

   $$y(x,t) = \hat{y} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} + \frac{t}{T}\right)\right].$$

   describes a wave. (Please mind the plus sign in the argument of the sine function.) What is the difference between this wave and that described by equation (4.1)?

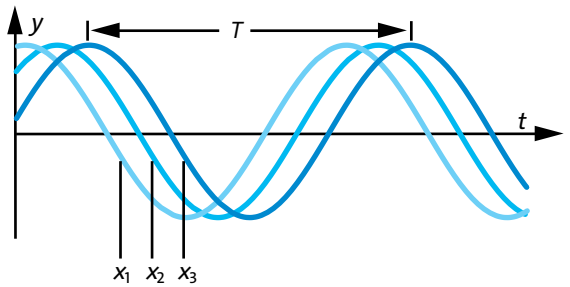

Fig. 4.9 The functions that correspond to the three positions x_1, x_2 and x_3 only differ from each other in the starting phase.

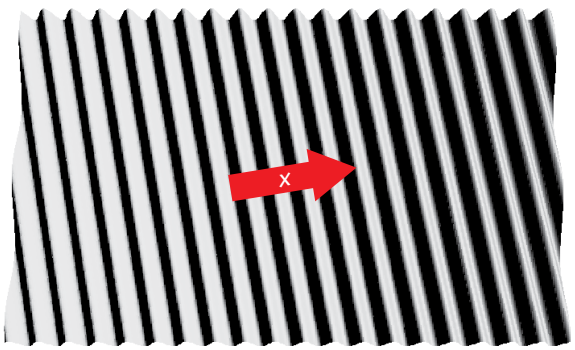

Fig. 4.10 Section of a sine wave on a two-dimensional carrier, for example on a water surface

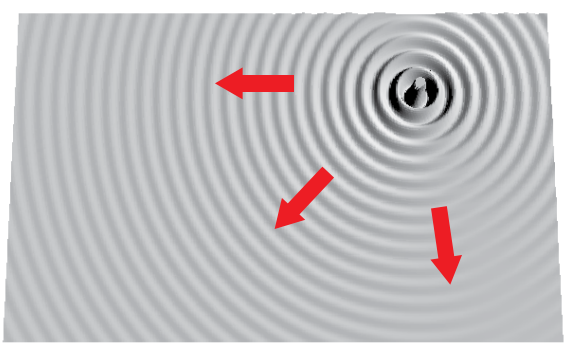

Fig. 4.11 Circular wave on a two-dimensional carrier

4.5 The relationship between velocity, frequency and wavelength

We examine how the wave crests and wave troughs come out of the sender. Per period T, exactly one complete wave crest plus one complete wave trough emerges: a wave piece with the length λ. This means that the whole wave moves forward by λ during the time interval T.

We therefore know the velocity of the wave. As the velocity is equal to the traveled distance divided by the time required for this distance, we obtain

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

We now replace the period T by the frequency f. With $T = 1/f$ we obtain

$$v = \lambda \cdot f.$$

Example: if a wave has the wavelength $\lambda = 2$ m and if every part of the wave carrier oscillates with the frequency 4 Hz, the wave will have the velocity:

$$v = 2\ \text{m} \cdot 4\ \text{Hz} = 2\ \text{m} \cdot 4/\text{s} = 8\ \text{m/s}.$$

Exercises

1. The velocity of sound waves in the air is approximately 300 m/s. What is the wavelength of the wave that corresponds to the standard pitch? The standard pitch has a frequency of 440 Hz.
2. Radio waves have a velocity of $v = 300\,000$ km/s. A sender sends with a frequency of 98.4 MHz. What is the wavelength of the waves?

4.6 Sound waves

In our examination of waves, we will always jump back and forth between the analysis of properties that all wave types have in common and special features of the individual wave types. After having gathered a lot of general properties, we will now take a detailed look at two particular wave types, first the sound waves and afterwards the electromagnetic waves.

Air is a carrier of the sound waves. As air is invisible, the sound waves cannot be seen either. (However, there is generally nothing visible of a sound wave moving through a solid material either. The amplitude is simply too small.) But the fact that sound must be a wave in the air can be seen well when looking at the formation of sound in a speaker. We need a speaker whose membrane is not covered, i.e. a speaker without its box.

Fig. 4.12 shows the design of a speaker. The membrane is suspended in a way that it can be displaced elastically and perpendicularly to the speaker. A coil is fastened at the rear of the membrane. The coil extends into a permanent magnet. One of the magnetic poles is located on the outer side of the internal part of the magnet, the other one on the inner sides of the external part. When an electric current flows through the coil, it is pushed forwards or backwards – depending on the direction of the electric current – by the magnetic field. Thereby, the membrane is moved as well.

We connect the speaker to a battery by means of a switch, Fig. 4.13. If we now close the switch, we will hear a cracking noise. And if we open the switch, there will be another cracking sound. When switching the system on, we can also see how the membrane jumps out of its initial resting position just as we can see it jump back during the switch-off.

The fast displacement of the membrane causes the pressure of the air, which moves directly in front of the membrane, to increase beyond or to fall below the normal pressure. This deviation from the normal state of the air detaches itself from the speaker and moves away. This moving deviation from the normal state is what we call a sound wave.

The air moves back and forth in the same direction in which the sound wave is moving. Hence, sound is a

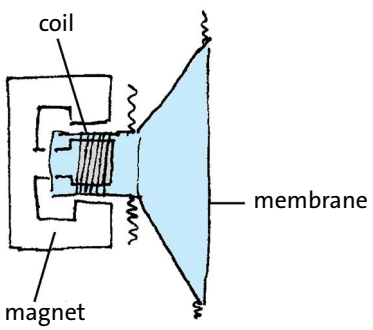

Fig. 4.12 Design of a speaker

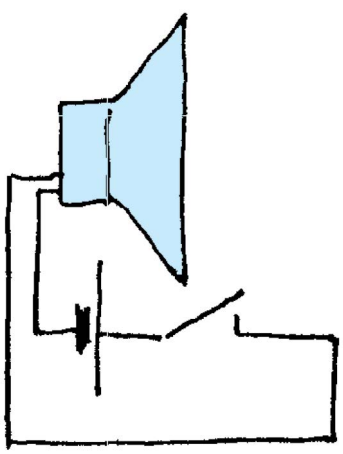

Fig. 4.13 We can hear a cracking noise during closing and opening of the switch.

longitudinal wave – at least in air. Later we will learn about acoustic transverse waves. Such transverse waves, however, can only exist in solid materials.

There is also another, simpler experiment that demonstrates that the carrier of the sound must be the air, Fig. 4.14. A bell is connected to a battery and placed under a glass dome. While the bell rings, air is being pumped out of the bell. During pumping, the sound becomes increasingly quiet until it can almost not be heard anymore. As soon as we let the air flow back in, its ringtone will become loud again.

The fact that we can still hear something in spite of the missing air is due to the sound being transported through solid substances as well, in this case the base of the bell.

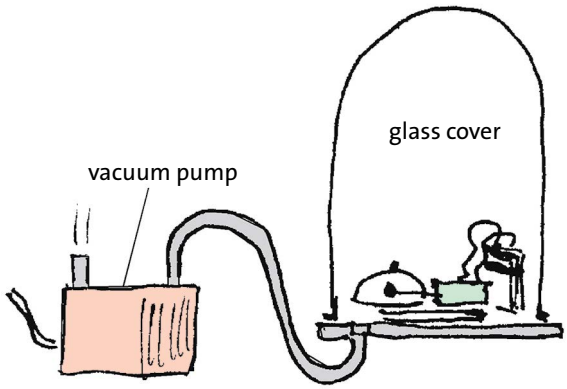

Fig. 4.14 When there is no air in the glass dome anymore, the bell can (almost) not be heard any longer.

> The air is usually the carrier of the sound waves. The sound (in the air) is a longitudinal wave.

We connect a loudspeaker to an electric energy source that switches the voltage periodically on and off. Fig. 4.15 shows this „squarewave voltage" as a function of time. The source should be designed in a way that the frequency can be changed. To begin with, we choose a very low frequency: approximately 1 Hz. We can hear two cracking sounds per second – one during each switch-on and one during each switch-off process.

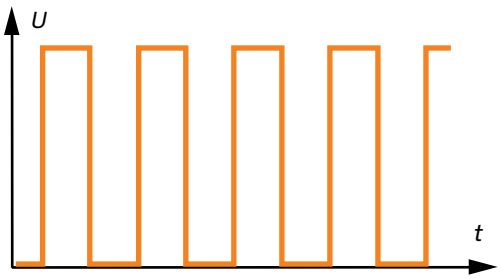

Fig. 4.15 „Square-wave voltage"

Now we slowly increase the frequency. The cracking noises come in an increasingly faster succession. Upon reaching approximately 20 Hz, our ear will no longer be able to perceive the cracking noises separately. What we hear is a *sound*: a lasting, constant sensation.

The sound that we perceive at first is a low-pitched sound. If we increase the frequency further, it will become increasingly highpitched.

Instead of a square-wave voltage, we now apply a sine voltage to the speaker. Again, we start with a very low frequency. But this time we do not hear anything below 20 Hz. Our auditory system is only sensitive for sine waves in a specific frequency range: approximately from 20 Hz to 20 000 Hz. As we become older, this range decreases. With a growing age, the upper limit moves towards lower frequency values.

> Our auditory system is sensitive for sine waves with frequencies from approximately 20 Hz to 20 kHz. The higher the frequency, the higher pitched the sound.

In a sine-shaped sound wave, both the velocity v of the air as well as the deviation Δp from the normal pressure behave according to equation (4.1). Hence, the following applies

$$v(x,t) = \hat{v} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right] \tag{4.2}$$

and

$$E(x,t) = \hat{E} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right] \tag{4.3}$$

The fact that the position coordinate x stands in the argument of the sine function, but not y and z, means that the wave moves in the x-direction. Thus the equations describe a plane sound wave.

We measure the velocity of a sound wave, Fig. 4.16. The wave is created by someone clapping his or her hands.

The wave passes the microphones M1 and M2 successively. These microphones are connected to an electronic stopwatch. When a signal comes from M1, the stopwatch starts running and will be stopped when a signal comes from M2. Hence, it measures the time that the wave needs to travel the distance from M1 to M2.

We obtain the velocity of the wave by dividing the distance between the microphones by the runtime. The measurement is not very accurate. With a more accurate measurement method we could find that the sound velocity depends on the temperature. Memorize 300 m/s as an approximate value.

> The velocity of sound waves in air is approximately 300 m/s.

The sound created by the speaker membrane is a longitudinal wave. We would like to create a transverse sound wave as a thought experiment. Instead of moving the membrane in the direction of its perpendicular axis, we move it back and forth in parallel to itself. You can imagine that there is no need to try this out. It cannot work because the membrane simply slides past the air instead of setting the air in motion. The situation is similar for liquids.

In gases and liquids only longitudinal sound waves can exist. Things are different in solid media. Although they are difficult to deform due to their hardness, every material has a certain elasticity and can therefore be deformed to some extent. A short impact against the end of a long bar leads to a slight deformation that runs through the bar, Fig. 4.17.

Earthquakes are an example of transverse waves in solid materials. From the place of origin of the earthquake, waves move away in all directions. Even at a distance of up to a hundred kilometers, they can cause destruction, Fig. 4.18. We can observe that seismic waves arrive in two phases: first the p-waves (p for primary) and then the s-waves (s for secondary). The p-waves are longitudinal waves while the s-waves are transverse waves (here is another way to remember this: p for push and s for shake.) The p-waves move approximately twice as fast as the s-waves.

Exercises

1. Name different sound sources, i.e. generators of sound waves.
2. What is the frequency of a sound wave with a wavelength of 2 m?
3. What are the wavelengths of the lowest pitched sound and the highest pitched sound that we are able to hear?
4. The sound velocity grows with an increasing air temperature. Assuming that a sine-shaped sound wave moves from an area of cold air to an area of warm air, what happens to its frequency and its wavelength?
5. During a thunderstorm you will see lightning and hear the thunder 10 seconds later. How far away from you is the thunderstorm?

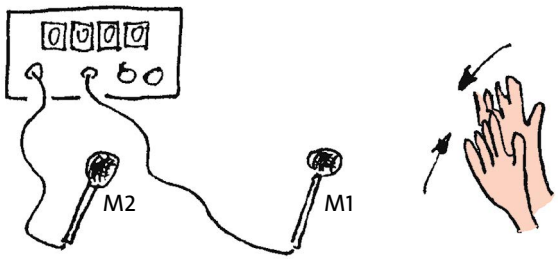

Fig. 4.16 Measuring of the velocity of a sound wave

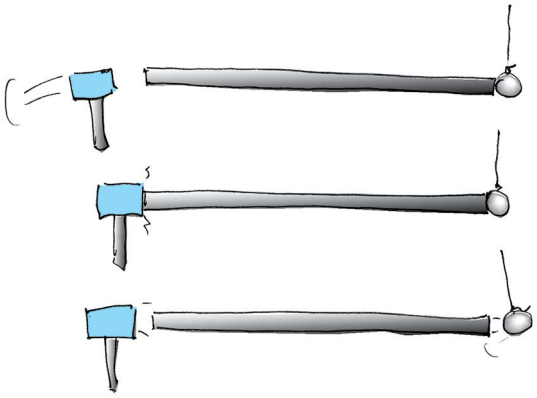

Fig. 4.17 A sound wave is created with a hammer. The arriving sound wave pushes the ball away.

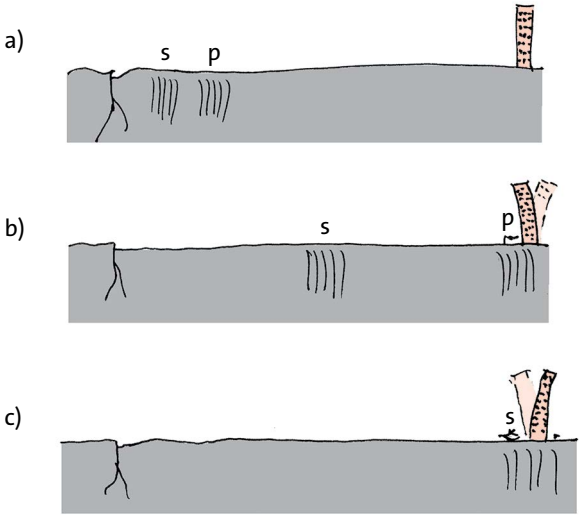

Fig. 4.18 (a) Formation of seismic waves. (b) A bit later the p-waves reach the skyscraper. (c) The s-waves arrive at the skyscraper even later.

4.7 Electromagnetic waves

We would like to create an electromagnetic wave. Therefore, we use the same method as for the creation of other waves: we have to ensure that the state of a carrier will change quickly at a certain point. As a sender, we use a wire in which an electric current can flow. While the current is flowing, the wire is surrounded by a magnetic field. We can therefore change the state of the environment of the wire: when no current is flowing, there is no magnetic field; when a current is flowing, a field exists.

By switching the current on and back off very quickly by means of closing and re-opening a switch within a short time, we can cause a fast change of the state of the environment of the wire. The changed state – i.e. the magnetic field – detaches itself from the wire and moves away by itself, Fig. 4.19. To understand how the magnetic field can move away, we have to remember that every time a magnetic field changes, an electric field emerges or disappears. Another magnetic field emerges or disappears through the change of the electric field. These interdependent processes constitute the wave. Hence, not only a magnetic field moves away from the wire, but also an electric field. Now you understand why these waves are called electromagnetic waves.

The wire from which the wave originates is in this case also called „transmitter antenna".

> If we switch an electric current in a wire on and off, an electromagnetic wave will run away from the wire.

Although the way in which we have described the formation of an electromagnetic wave works in principle, its practical feasibility is very poor because it is difficult to switch an electric current on and off fast enough as to create an observable wave. If we use a common switch, the switching process will be much too slow for our purposes: when touching the contacts for the first time, the current starts rising relatively „slowly". Also, the current strength does not decrease as fast as we would require when the switch is opened. A better method is to create a spark discharge by means of a very high voltage. Such a spark discharge comes with a much stronger change of the electric current.

In principle, an electromagnetic wave could also be created in a different way: taking a permanent magnet and moving it back and forth very quickly. Also in this case, the magnetic field changes. This method, however, is not working well because we will not be able to move the magnet fast enough.

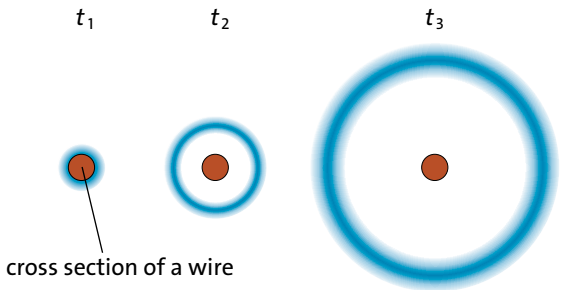

Fig. 4.19 The wire and the wave (in crosssection) at three different instants of times. The electric current is switched on and off very quickly.

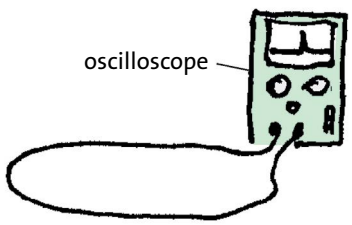

Fig. 4.20 The wire loop is the receiving antenna. In this antenna, a voltage is induced by the arriving wave. The oscilloscope indicates this voltage.

But how can we tell whether a wave has been created at all? We need a device that reacts to the arrival of the wave: a „receiver antenna". The simplest method: we switch on a radio and set it to short, medium or long wave (only FM is not convenient for our purpose). Every time one of the waves created by us arrives, we hear a cracking noise. The arrival of the wave can also be made visible. We connect the two ends of a wire with the input of an oscilloscope, Fig. 4.20. The wire forms a coil with a single winding. The arrival of the wave means a change of the magnetic field in this coil. This magnetic field change causes a voltage between the ends of the coil. We have called this process electromagnetic induction. As the change happens within a very short time, only a very short „voltage pulse" is produced. This pulse can be seen on the screen of the oscilloscope.

One important question is still unanswered: which is the medium in which the electromagnetic wave is moving? What is the carrier of the wave? It cannot be the air: electromagnetic waves also move through matter-free spaces. The light, which is indeed an electromagnetic wave, passes the 150 million km from the Sun to the Earth without any problem, i.e. through a space that is practically free of air and any other mat-

ter. (The atmosphere of the Earth only has a thickness of several km.)

We can therefore conclude that the so-called empty space has to contain something that assumes the role of the carrier of the electromagnetic waves. When it was discovered that light is a wave, this „something" was called „ether". At first, scientists believed that light was a mechanical wave in this ether, i.e. a wave whose carrier is moving just as the air moves in case of sound waves. Only later, they found that the change of state of the carrier of the electromagnetic waves is not a deformation and that this carrier has other surprising characteristics as well.

Consequently, it was renamed because too many obsolete concepts had been associated with the name „ether". This new name is „vacuum", meaning „emptiness".

> The carrier of the electromagnetic waves is called „vacuum".

We must not misunderstand this name. Emptiness is not the same as „nothing". Nothing means that there is no substance at all. In an empty recipient, by contrast, there can still be something else. An empty coke bottle contains no more coke but there is generally air in it. In an empty gum vending machine, there are no more chewing gums but there is still the entire vending machine mechanism. And although we cannot take any more electric energy out of an empty battery, the battery is still full of lead sulphate and sulphuric acid.

When we say that there is a vacuum in a region of space, we mean that, although there is no matter in the chemical sense, there can still be something else: precisely the carrier of the electromagnetic wave. As long as no wave moves through the vacuum, the vacuum is in its „ground state".

The electromagnetic waves that we created were short pulses, similar to the few individual sound wave pulses in the air that we had created with the speaker at first. To create a permanent wave, we have to switch the electric current in our transmitting antenna on and off in quick succession. And if we want to have an electromagnetic sine wave, we will have to send a sine-shaped alternating current through the wire. To make the antenna emit a wave, however, the frequency has to be very high. Only an extremely weak wave will be formed with the 50 Hz of the normal alternating current.

In a sine wave, both the electric field strength E as well as the magnetic field strength H behave in accordance with equation (4.1).

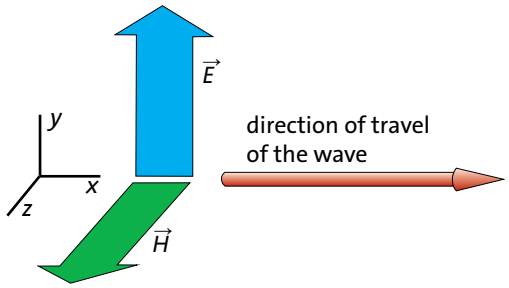

Fig. 4.21 Electric field strength vector, magnetic field strength vector and direction of propagation of an electromagnetic wave

Hence, we have:

$$E(x,t) = \hat{E} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right] \tag{4.4}$$

and

$$H(x,t) = \hat{H} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right]. \tag{4.5}$$

Also in this context, the fact that the position coordinate x, but not y and z, stands in the argument of the sine function means that the wave moves in the x-direction. The equations describe a plane electromagnetic wave.

Both the direction of the electric as well as of the magnetic field strength vector is orthogonal to the x-direction. In addition, the electric and the magnetic field strength vector are orthogonal to each other. When we position the y-axis in the direction of the electric field strength, the magnetic field strength vector points in the z-direction, Fig. 4.21.

Bear in mind that the electric and the magnetic field strength are „in phase": when the electric field strength takes on a maximum value, also the magnetic field strength has a maximum value.

You already know the velocity of movement of electromagnetic waves. In vacuum it is 300 000 km/s.

Electromagnetic waves with a variety of wavelengths surround us at all times. On one hand, there are natural sources of electromagnetic waves with diverse wavelengths. On the other hand, waves of many different wavelengths are also generated and used technically.

The wavelength range of the waves that we create or observe reaches from a millionth of a nanometer up to kilometers. Although these waves are all of the same nature and although they only differ from each other in their wavelength, the creation methods are very different. In addition, many things happen when waves of diverse wavelength ranges hit matter. This is the reason

why they can be used for a variety of purposes and why the waves were given different names depending on the wavelength range: gamma radiation, X-rays, ultraviolet radiation (or ultraviolet light), („visible“) light, infrared radiation (or infrared light), microwaves, radio waves.

Exercises

1. Why does a thunderstorm interfere with radio reception?
2. Name different sources of electromagnetic waves.
3. How can we tell from equations (4.4) and (4.5) that the electric and the magnetic field strength are in phase?
4. Draw a three-dimensional coordinate system of the position coordinates x, y and z in perspective. Sketch the electric field lines (in one color) and the magnetic field lines (in a different color) for a defined time (snapshot).

4.8 Energy transport with waves

Waves transport energy. We would like to examine the energy transport of sine waves as an example. First, we will formulate the question more clearly.

The energy transport is described by the physical quantity P, i.e. the energy current strength or energy current in short. This quantity tells us how much energy flows through a chosen area per second. No energy at all flows through an area that is parallel to the direction of propagation of the wave. We therefore choose the area that is orthogonal to the direction of propagation, i.e. parallel to the wave fronts, Fig. 4.22.

But the energy current depends of course on the size of the area. When looking for a measure for the energy flow that is independent of the area, we have to divide by the surface area. This is how we obtain the energy current per area, i.e. the *energy current density* j_E:

$$j_E = \frac{P}{A}$$

Energy current density = energy current divided by the area

As the wave is a sine wave, the energy current density will change at every point in the rhythm of the passing wave.

Although the energy current density that changes over time can be calculated, we will be more interested in its time average value in most cases. Therefore, we like to know:

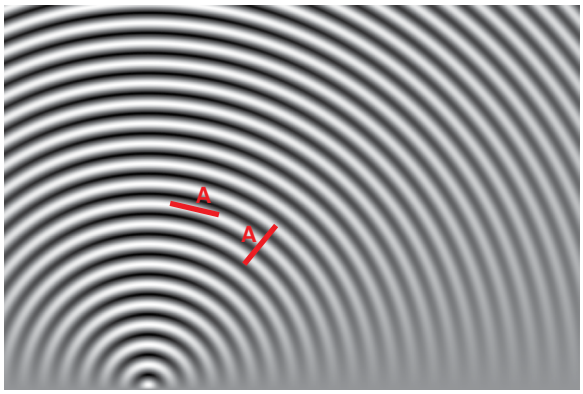

Fig. 4.22 No energy flows through the area A that is parallel to the movement direction of the wave (right). The energy current reaches its maximum when the area is parallel to the wave fronts (left).

$\overline{j_E}$ = time average of j_E.

We look for the relationship between this average *energy current density* and the other quantities that we use to describe the wave. But which other quantities? We can describe the same wave with different quantities: sound waves with the velocity or with the pressure, equations (4.2) and (4.3), electromagnetic waves with the electric or the magnetic field strength, equations (4.4) and (4.5). And there are even other possibilities. Each of these quantities, however, behaves in accordance with an equation of the same form: equation (4.1) that we repeat once again at this point:

$$y(x,t) = \hat{y} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right]. \qquad (4.6)$$

Regardless of the quantities that we look at, the essential features of the relationship with j_E are the same. In the following, you can imagine that we use y to describe the movement of the air in a sound wave. But the result will also apply for the other quantities.

Instead of making long and complex calculations, we derive the relationship by means of skilled guessing. At first, we check the simplest assumption that we might think of: the time average value of j_E is proportional to the time average value of y, i.e.:

$$\overline{j_E} \sim \overline{y(x,t)} = \overline{\hat{y} \cdot \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right]}.$$

However, this relationship cannot be correct as the temporal time average of the sine term on the right side of the equation is zero. The average energy current density would consequently be zero, too, which is certainly not the case. A slightly more complicated as-

sumption would be that the j_E average value is proportional to the average value of the square of y. As you know, the square of a number is always positive. Hence, the following would apply

$$\overline{j_E} \sim \overline{y(x,t)} = \hat{y}^2 \cdot \overline{\left\{\sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right]\right\}^2}. \qquad (4.7)$$

The time average value of the bracket on the right side of the equation is 1/2.

Therefore, we would have

$$\overline{j_E} \sim \frac{1}{2}\hat{y}^2.$$

According to this assumption, the average energy current density would be proportional to the square of the amplitude – and this is actually correct. The quantities that we use to describe the wave are not relevant in this context: the pressure or the velocity for the sound wave, the electric or the magnetic field strength for the electromagnetic wave. The average energy current density tells us how much energy is transported by the wave and therefore also how „vigorous" the wave is at a certain point, i.e. the strength of the movement in case of a mechanical wave and the strength of the fields in case of an electromagnetic wave. As we will often refer to this characteristic in the following, we introduce an abbreviated name for the long denomination „time average of the energy current density": *intensity*.

> The intensity of a sine wave is proportional to the square of the amplitude.

Exercises

1. Not only the energy current is proportional to the square of quantities that can admit positive and negative values, but also the energy content. Give four examples. If you do not remember the formulas, look up in your physics text book. This is also a good opportunity to memorize them.
2. Why is the time average value of the sine term on the right side of equation (4.7) equal to 1/2?

4.9 Two waves at the same place

We examine what happens when two waves collide with each other. Will there be a clash?

We can see it best in the case of waves on a string. We simultaneously send out a wave from each of the two ends of a long string that is lying on the floor, Fig. 4.23. The waves move towards each other and then other two waves move away in both directions from the meeting point. What happened? Did the two waves rebound from one another. Were they reflected on each other?

We change the experiment slightly. From one end of the string we send out a deviation to one side (transversal to the string) and from the other end a deviation to the opposite side, Fig. 4.24. The waves that arrive at the ends after the collision are those that had been sent out from the opposite end. Hence, the waves have not been reflected on one another but they have „moved through each other". One wave is not changed by the other one and each of the two waves moves as though the other one was not there.

Here is what we could not see in our experiments as things were going too fast:

When the waves meet halfway, the deviations are added up. For the waves from Fig. 4.23, this means that a single wave with the double deviation is located in the middle of the string for a moment, Fig. 4.25.

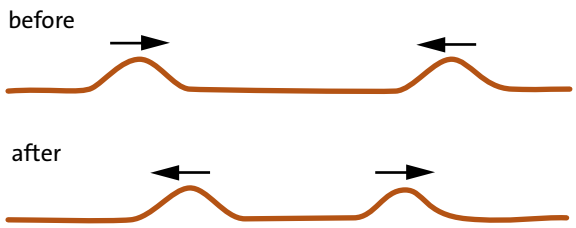

Fig. 4.23 Were the waves reflected on each other?

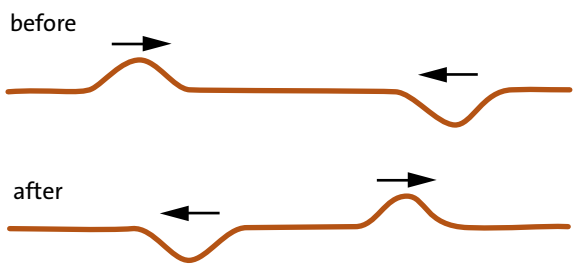

Fig. 4.24 The waves move through each other.

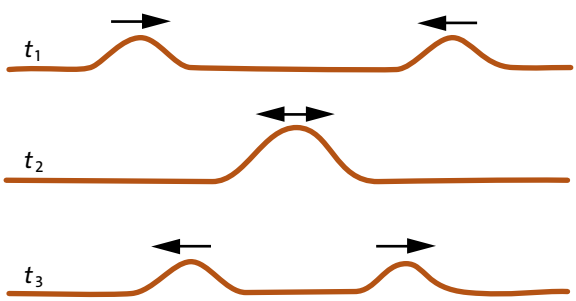

Fig. 4.25 At the time t_2, there is a wave crest with the double deviation at the center.

And in the case of the waves from Fig. 4.24, the string is completely straight for a moment. The deviations add up to zero, Fig. 4.26. Of course, this addition rule already applies prior to the collision of the waves and also after the waves have moved away from each other. However, the deviation of one wave will then be zero at the place of the other wave. It is consequently not surprising that the waves can move through each other in an undisturbed way.

The addition rule does not only apply for the deviations of a wave on a string, but also for all the other quantities that we had denominated with y: for the velocity and the pressure in the case of sound waves and for the electric and the magnetic field strength in the case of electromagnetic waves.

These somehow unhandy formulations can be summarized as follows:

> Waves move undisturbed through each other.

Just like some other rules, this rule does not always apply. It will no longer be valid if the deviations of the waves are too large. Example: two large waves in the sea that are about to break over and that move towards each other will no longer move through each other undisturbed. Most waves that we come across, also sound and electromagnetic waves, however, are so weak that our statement is fulfilled very well.

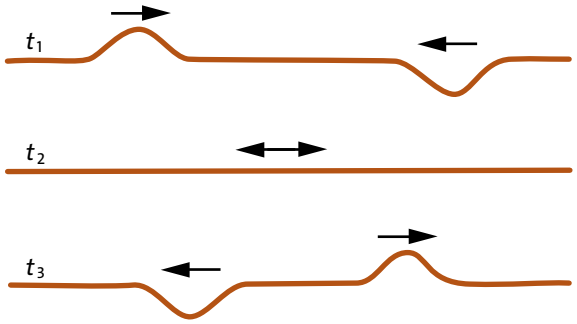

Fig. 4.26 At the time t_2, the string is completely straight.

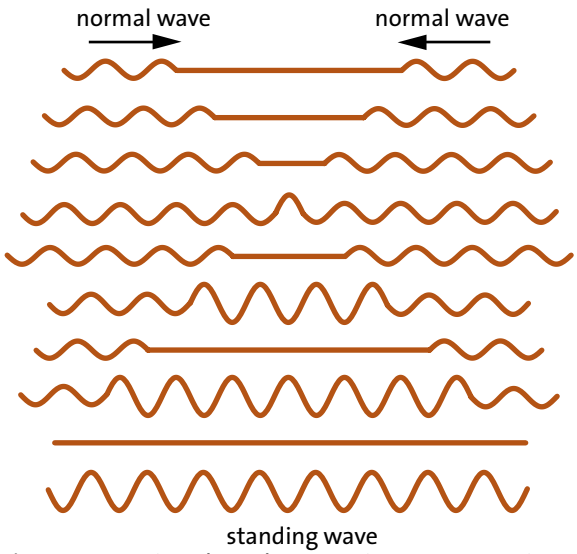

Fig. 4.27 Two sine-shaped wave trains move opposite to each other and through each other. The process is displayed for ten successive points in time.

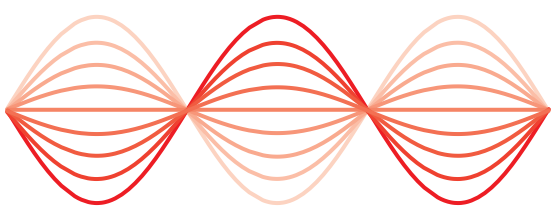

Fig. 4.28 The different hues correspond to the standing wave at nine different points in time.

Exercises

1. Is wind a wave? Can „ two winds“ move through each other in an undisturbed way?
2. When two electromagnetic waves move through each other, the overall field strengths can be calculated by adding up the field strengths of the two individual waves. We have already come across a similar matter earlier as we discussed simple electric and magnetic fields. What was the rule back then?

4.10 Two sine waves – interference

We experiment once again with the string: we stretch it out on the floor and start sending out sine waves on each of the two ends, Fig. 4.27. The wave trains move towards each other until they meet and subsequently move through each other.

What we see is quite peculiar. There is no longer a movement in one or in the other direction of the string. Fig. 4.28 shows a section of the string at nine different instants of time. The string is sine-shaped, but the height of the crests and troughs changes. The points where the string is deviated neither in one nor in the other string direction remain fixed, i.e. they do not move in one or another string direction as would be the case for a normal wave.

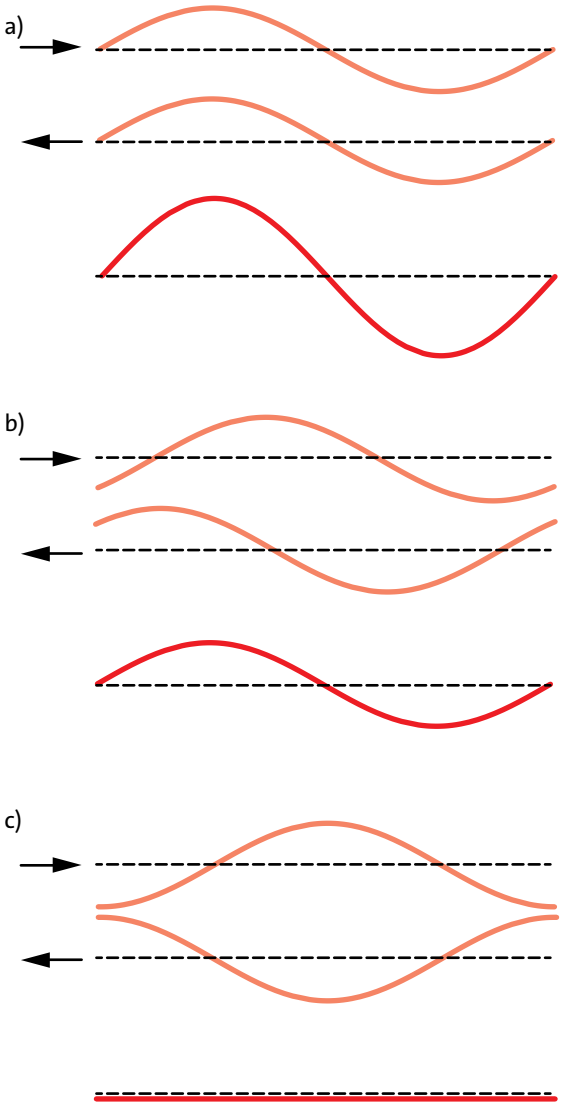

Fig. 4.29 The standing wave is formed by adding up the deviations of two waves moving opposite to each other. The image sections a, b and c show the addition for three points in time.

Such a phenomenon is called *standing wave*. The points of the strongest movement are the *antinodes*; the points of the string that do not move are the *nodes*.

Fig. 4.29 shows how the nodes and antinodes are formed. Please bear in mind that the standing wave is a superposition of two sine waves moving opposite to each other. The deviation of the standing wave can be obtained by adding up the deviations of the two wave parts.

For three different instants of time each Figure a, b and c shows at the top of the two wave components, and below it displays the wave that actually arises through addition. We can see that in the antinodes, the amplitude of the standing wave is twice as large as in the component waves: here, two deviations of the same direction come together each time. At the places of the nodes, the deviations of the component wave are opposite. Their sum is zero.

We can also see that the distance between two neighboring nodes is half a wavelength.

Also the formation of a standing wave can be observed much better by looking at the wave function as a computer animation. We enter:

$$y = \sin(x - t) + \sin(x + t).$$

The first addend on the right side describes a wave moving to the right; the second addend describes a wave moving to the left. Hence, the sum represents two waves moving in opposite directions.

For a standing wave to be formed, the amplitudes of the waves moving opposite to each other have to be the same. But amplification and attenuation will also occur if the component waves do not have the same amplitude.

The process in which the superposition of two waves leads to amplification at some places and to attenuation or extinction at other places is called *interference*. We also say that the two waves interfere with one another.

> A *standing wave* arises if two sine waves with the same amplitude and wavelength move in opposite directions.
>
> The distance between two adjacent nodes is half of the wavelength.
>
> The process of mutual amplification and attenuation of waves is called *interference*.

Exercises

1. What happens when two sine waves with the same wavelength but different amplitudes move through each other after coming from opposite directions?
2. What happens when two sine waves, which have the same amplitude and wavelength and which move in the same direction, superpose?
3. In case you have a suitable algebra program: let the function $y = \sin(x - t) + \sin(x + t)$ run as an animation. Change the amplitude of one of the two waves, i.e. for example: $y = \sin(x - t) + (2 \cdot \sin(x + t))$. Change the frequency of one of the two waves, i.e. for example: $y = \sin(x - t) + \sin(x + 2t)$. Describe the results.

4.11 Reflection of waves

A convenient method to create standing waves is to let a sine wave be reflected. The reflected wave superposes with the incoming one whereby a standing wave is formed. The method works for example with waves on a string. End A of the string is attached somewhere. End B is moved back and forth in a sine-shaped way. A sine wave starts moving from B and is reflected at A. The result is a standing wave. But this standing wave is only well-shaped near the end A because only there, the two wave parts have the same amplitude. At B, the wave moving in the direction A is much stronger than the wave coming from A because something of each wave is lost on the way.

Our method to create standing waves also works with sound waves, Fig. 4.30. The speaker sends a sine wave against the wall where the wave is reflected. The reflected and the incoming wave move through each other and a standing sound wave results.

We move a microphone that is connected to an oscilloscope between the speaker and the wall. We can clearly see the points of amplification and attenuation on the oscilloscope. As sound waves are three-dimensional waves, nodes and antinodes are surfaces.

Electromagnetic waves are reflected on electrically conductive surfaces. If we let an electromagnetic wave move perpendicularly against a metal wall, a standing electromagnetic wave will be formed. The shorter the wavelength, the smoother the „reflector" has to be. A very smooth metal surface is needed for the reflection of light: a mirror. If we send monochromatic light (i.e. a sine-shaped light wave) onto a mirror, a standing light wave will be formed in front of the mirror. As the wavelength of the light is very short, the nodes, however, are so close to one another that it is difficult to detect them. Later we will get to know a trick that enables us to increase the distances between the nodes.

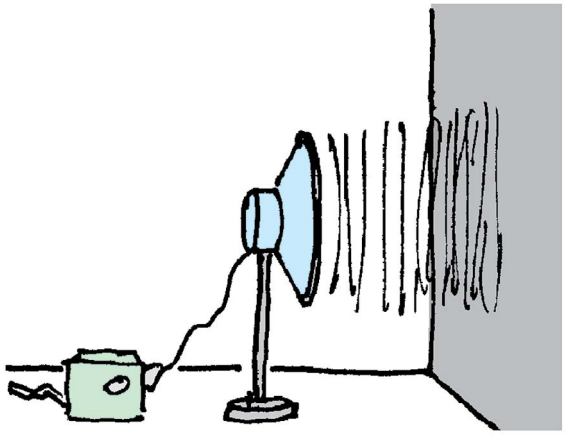

Fig. 4.30 Creation of standing sound waves by reflection on a wall

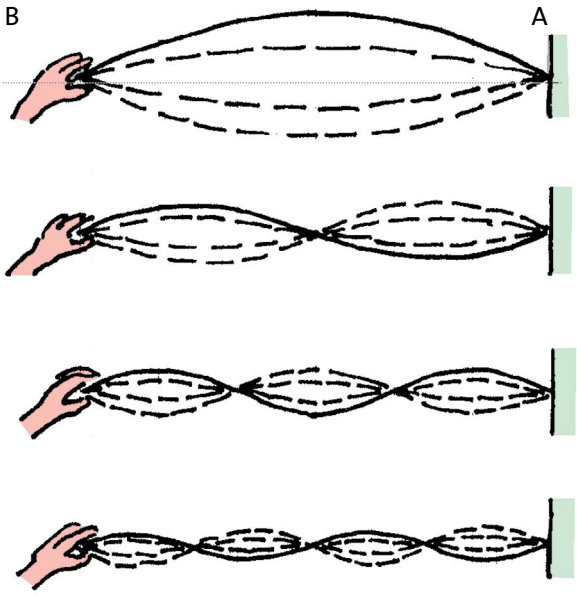

Fig. 4.31 Natural oscillations of a string. The length of the string can accommodate an integer multiple of half the wavelength.

4.12 Natural oscillations of wave carriers

Again, we attach the end A of our string but this time we hold end B in the air and tighten the string, Fig. 4.31.

It will not touch the ground during the whole experiment. Therefore, friction is lower than in the previous experiments and the waves in the string are attenuated much less. As we create a sine wave here, this wave will not only be reflected at A but it will move back and be reflected a second time at B and subsequently a third time at A, and so forth. If the returning wave crests arrive at B in the right moment, there will be an amplification: the deviation of the wave reflected by A is added to the deviation of the newly created wave at B in a way that an amplification occurs. The movement of the string „builds up" to a standing wave that has a node both at A and at B. For such a wave to

arise, the string length l has to match exactly with an integer multiple of half of the wavelength. Thus we have:

$$l = n \cdot \lambda/2 \qquad n = 1, 2, 3... \tag{4.8}$$

As each wavelength is associated with a specific frequency, the string end has to be moved with a very specific frequency to obtain one of the possible standing waves.

You have probably noticed by now that we have already discussed this phenomenon earlier, in sections 3.4 and 3.5. We are talking about *natural oscillations* of the string. Now we can see that such natural oscillations can be interpreted as two sine waves moving in opposite directions.

> A natural oscillation can be considered as two sine waves moving in opposite directions. The following applies:
> Amplitudes and wavelengths are equal.
> The wavelength is in line with condition (4.8).

We have explained the standing waves with the example of waves on a string because they illustrate the processes particularly well: the wave carrier is visible and the movements are pretty slow. But standing waves can also be created in or on any other wave carrier. This requires two reflectors standing opposite to one another. Depending on the type of the wave, something else is used for reflection, Table 4.1.

Wave type	Wave is reflected on
wave on a string	fastening devices of the string
water waves, sound waves	smooth walls
sound waves in a tube	(closed or open) tube end
electromagnetic waves	electrically conductive surface, mirrors

Table 4.1 About the reflection of waves

The oscillations of the string of a musical instrument can be imagined as two „waves on a string" moving opposite to each other, and the oscillations of the air in a wind instrument may be imagined to consist of two sound waves moving opposite to each other. Notice that both an open as well as a closed tube reflects the wave. Equation (4.8) only applies when either both ends are open or when both ends are closed. In fact, some wind instruments are open on both ends, for example the flute, and some only on one end, for instance the clarinet and the brass instruments.

Up to now, we have clarified the conditions under which a standing wave can exist. We know how the oscillator that performs the natural oscillations must look like. What we have not yet fully clarified is the question of how to excite a natural oscillation. In principle, it works similar as for the simple oscillations that we had studied before. Also in this context, there are two methods: either using an „exciter", i.e. an energy source that already creates an oscillation with the right frequency itself, or ensuring an energy supply that is controlled by the natural oscillation itself. Again, the self-controlled oscillations are the most interesting ones. Just as in case of common oscillations, the following applies:

> To maintain a natural oscillation, we need
> - an oscillator (wave carrier + reflectors)
> - an energy source
> - a control of the energy supply.

Like in the case of simple oscillations, diverse and often quite complicated technical tricks are applied for the control. We will look at a few examples. However, we only ask about the oscillator and the energy source and not about the working principle of the self-control.

Musical instruments

In string instruments, the string performs a natural oscillation. It constantly loses energy to the body of the instrument, which, similar to a speaker membrane, emits a sound wave. When we pluck the string only slightly, it performs a strongly damped oscillation. Its energy is used up quickly. But when we strike the string with the bow, we constantly supply new energy. In fact, we would have to say: the string itself takes up the energy by jumping over the bow in sync with the oscillation.

In woodwind instruments, the air inside the instrument oscillates. The energy supply from the air current is controlled by the oscillating air inside the instrument.

In the case of harmonica, squeeze box, accordion and bandoneon, metal tabs oscillate in a similar way as the ruler from Fig. 1.2. Also in this case, the metal tab itself controls its energy supply from the air current.

Laser

In a laser, a standing light wave is created between two mirrors, Fig. 4.32.

Between the mirrors, there is a material whose electrons can be brought to an excited state by means of an electric current. As an excited state, we choose a state from which the electrons will not jump back to the ground state by themselves. The electrons are consequently charged with energy by the electric current, similar to the weight of the pendulum clock when it is pulled up. The natural oscillation of the light between the mirrors now takes up energy from this storage system. It makes the electrons emit light in sync with the natural oscillation. This process is called *stimulated emission*. (The word „laser“ is an acronym for „light amplification by stimulated emission of radiation“.)

To take advantage of the laser, one of the two mirrors is slightly transparent (approximately 5 %) so that some light can escape the laser. The energy that this light carries away has to be replaced repeatedly by means of stimulated emission.

Exercises

1. Generate standing water waves at home in your kitchen. You need a rectangular baking mold and a small cutting board that is a bit narrower than the baking mold. Fill around three quarters of the baking mold with water. Dip one end of the board perpendicularly in the water on one end of the baking mold. Move it up and down very slowly at first, without touching the walls of the baking mold. Almost nothing happens in the water. The water surface rises and drops but to such a slight extent that we can hardly see it. Now make the up-and-down movement a bit faster. The water will move increasingly vigorously. If you reach the correct frequency of the movement, the water will slop over. Now you have triggered the natural oscillation of the water. Describe the movement of the water. Now the first harmonic: move the board up and down in the middle of the baking mold. The frequency has to be higher now. Also here, we can make the water slop over. Describe the movement of the water once again.
2. An elastic thread with a length of 1 m is attached firmly on one end. Through continuous up-and-down movements of the free end, we can create a wave that moves through the thread at a velocity of 6 m/s and that is reflected on the ends. (a) What is the maximum wavelength that the created wave can have so that a standing wave is formed in the thread? (b) At which frequency does the free end have to be moved up and down so that a standing wave with two nodes is formed in the thread (in addition to the nodes on the thread ends)? (c) Sketch the movement of the thread in this case.
3. Find out how the self-control of the natural oscillation works in string instruments, in the recorder and in the harmonica (use the Internet, specialist literature, encyclopedias).

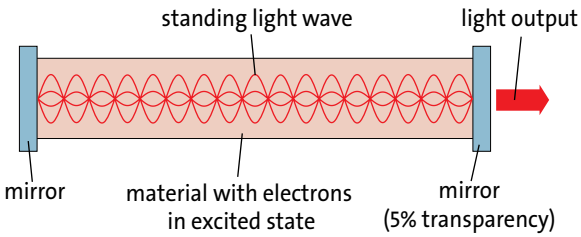

Fig. 4.32 Schematic illustration of the laser: a standing light wave is located between the mirrors. Some light continuously escapes through the right mirror. The respective quantity is supplied anew by the excited electrons of the laser material.

4.13 The interference of waves

We have already come across this phenomenon. When two sine waves with the same amplitude and the same wavelength move opposite to each other, the quantity y (i.e. the quantity that we use to describe the wave) is always zero at some points. At other points, its value changes in a sine-like way with an amplitude that is larger than that of the individual waves.

We now examine the interference for the case in which the two sine waves do no longer move exactly opposite to each other, but diagonally to one another. The Figures show waves on a two-dimensional carrier in a top view. But we can also imagine the pictures as sections through a three-dimensional wave carrier. Black stands for negative y-values, white for positive ones. The gray color outside of the wave areas characterizes $y = 0$.

Fig. 4.33 shows a snapshot of two waves that move through each other at an angle of 40°. The movement of the wave fronts is indicated by the arrows.

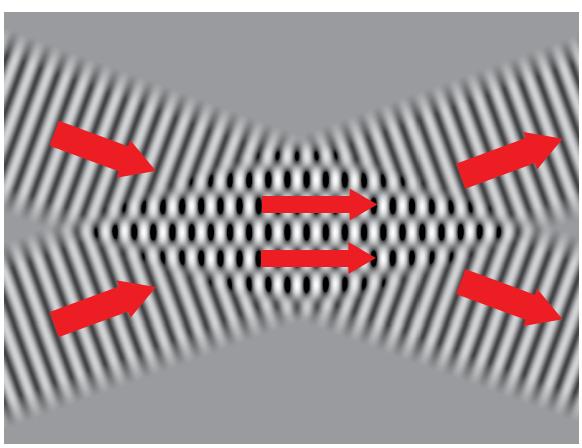

Fig. 4.33 Snapshot of two waves that intersect at an angle of 40°

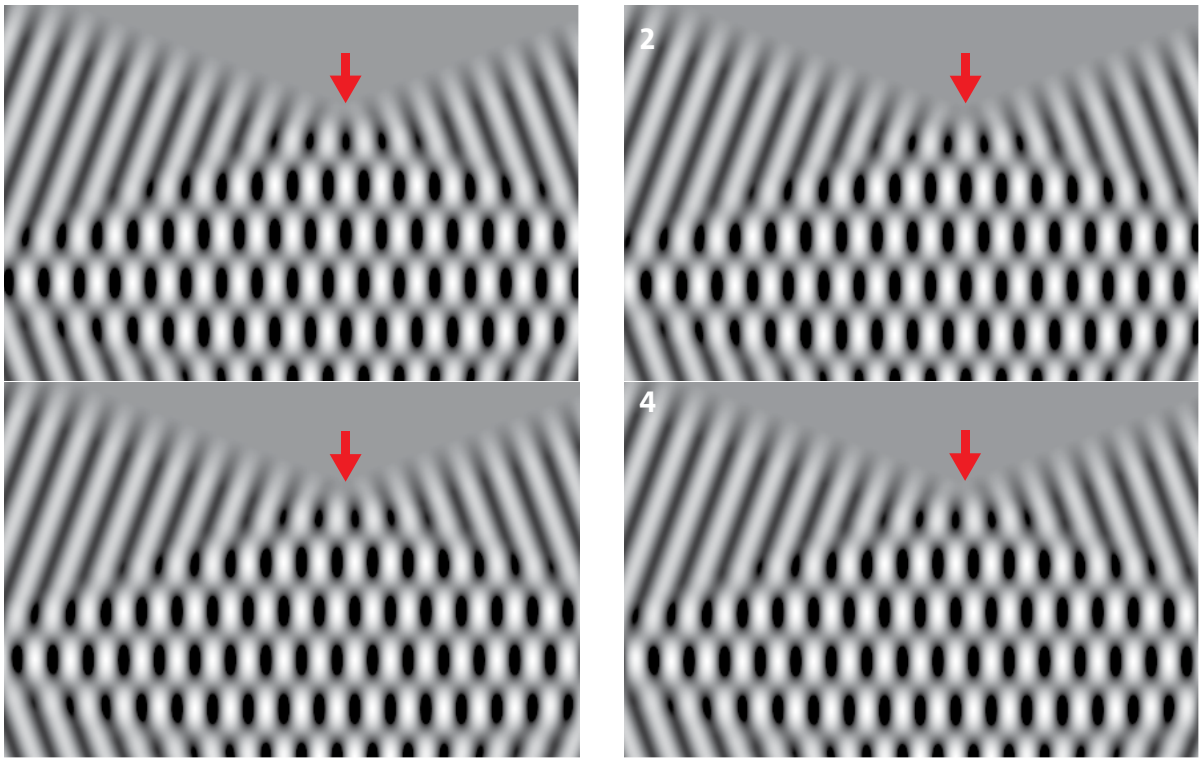

Fig. 4.34 From one image to the next, the wave advances by a quarter of a wavelength.

Again, it would be more convenient to see the wave as an animation. Fig. 4.34 is a somehow poor substitute for it. The picture shows the interfering waves at four different instants of time. From one picture to the next, the waves have advanced by a quarter of a wavelength. We can see the difference between the images if we look exactly at the place where the arrow points to. The arrow does not move along with the waves.

Here is the interesting aspect of these pictures: in the area that is covered by both waves, there are horizontal lines on which $y = 0$ at all times. At these points, the waves extinguish each other. In the middle between these straight lines, the y-values perform sine oscillations with an amplitude that is larger than that of the individual waves. Here, the two waves are amplified.

There is another method to explain the phenomenon: Fig. 4.35 shows a snapshot of the same wave as in Fig. 4.33 with the only difference that not y but y^2 is shown. Here we have

- white: $y^2 = 0$,
- black: y^2 = maximum.

If the picture were animated, the black areas would move as indicated by the arrows. If we make many of such snapshots and if we calculate the temporal mean value of y^2 at each point, we will obtain the intensity of the wave, Fig. 4.36. In this illustration, the interference can be seen best. In the white areas, the intensity is zero, and consequently also the temporal mean value of y^2 is zero. The intensity is highest at the black regions. Outside of the intersection area, it is equal everywhere. Fig. 4.36 shows that the energy that arrives with the two waves from the left is channeled in the intersection area by the stripes that are illustrated with a dark color in the Figure.

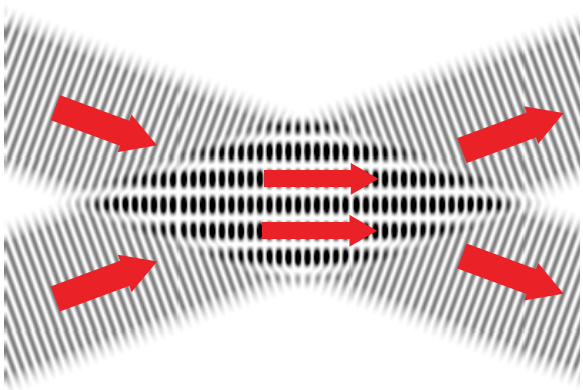

Fig. 4.35 Similar to Fig. 4.33, but the square of y is displayed.

Fig. 4.36 Similar to Fig. 4.35, but the temporal mean value of the *y* square is displayed here. We can see the interference in the intersection area of the two waves. Light stripes: extinction; dark stripes: amplification

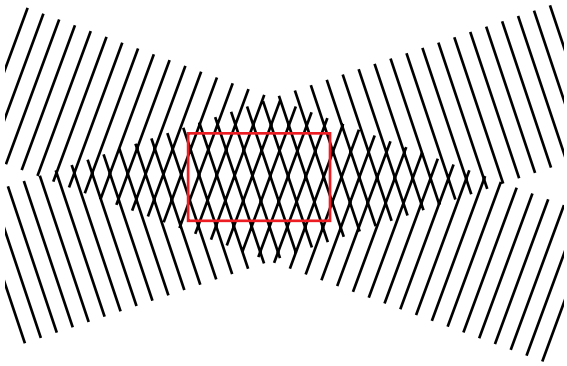

Fig. 4.37 The lines are the places of the maxima of the individual waves. The area of the frame is shown in an enlarged display in Fig. 4.38.

We would now like to calculate the distance between the stripes. Therefore, the process is once again illustrated schematically in Fig. 4.37. The lines indicate the position of the wave maxima at a specific instant.

The section marked by the black frame is displayed in an enlarged way in Fig. 4.38. Besides the maxima, also the minima are indicated by straight dashed lines in that picture. Amplification occurs where a maximum meets a maximum and where a minimum meets a minimum. Extinction occurs where a maximum meets a minimum. Both the points with amplification as well as those with extinction are located on straight lines as we have already concluded from the previous pictures.

Fig. 4.39 shows how we can calculate the distance a between two neighboring amplification lines.

The distance between two successive maxima in every individual wave is equal to the wavelength λ. We can therefore conclude from the Figure:

$$\sin\frac{\alpha}{2} = \frac{\lambda}{2a}. \qquad (4.9)$$

Here, α is the angle between the wave fronts of the two waves. We therefore obtain for the distance between the amplification lines:

$$a = \frac{\lambda}{2 \cdot \sin\frac{\alpha}{2}}.$$

We can read from the equation: a is greater than or equal to $\lambda/2$ because the values of the sine function are between 0 and 1. The smaller the angle α between the wave fronts, the larger is a. If we make α small enough,

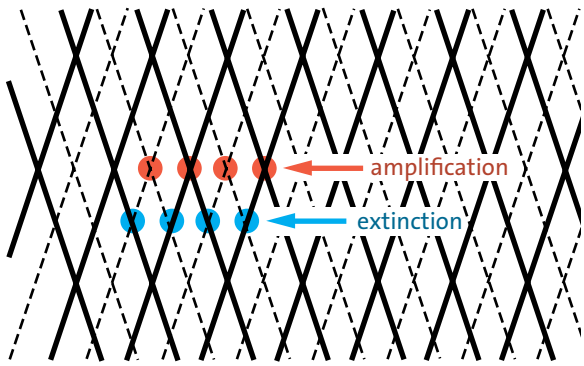

Fig. 4.38 Bold, continuous lines: wave maxima; thin, dashed lines: wave minima. The points where maxima meet and where minima meet are located respectively on a straight line, and so are the places where a maximum meets a minimum.

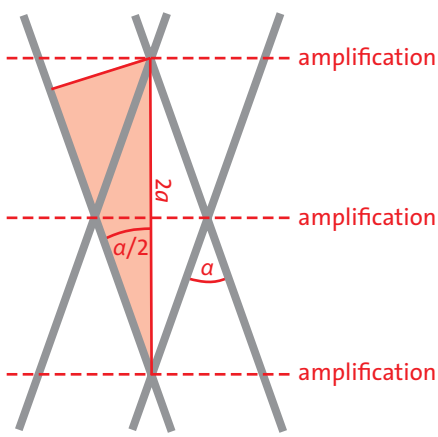

Fig. 4.39 *a* is the distance between two adjacent amplification lines.

the distance between the amplification areas can reach any size, Fig. 4.40.

> The smaller the angle between two interfering sine waves, the greater the distance between the amplification and the extinction areas.

Now we can see a first benefit of our discussions. The wavelength of light is so short that it cannot be measured with conventional methods such as using a ruler. But if we make two sine-shaped light waves intersect at a very small angle, we can obtain amplification and extinction areas with a distance that is long enough to be measured easily. A white screen is put in the way of the two light waves, i.e. placed in the interference area or, in other words, where the waves move through one another, Fig. 4.41. We can then see, with our naked eye, places of amplification and points of extinction on the shield. As the light waves also have an extension in a direction that is perpendicular to the drawing plane, light and dark stripes are formed on the screen (perpendicularly to the drawing plane): an *interference pattern*.

We can therefore measure the distance a. As we also know the angle α, we can calculate the wavelength of the light. This method is very important, but there is a snag: it is relatively difficult to produce sine-shaped light waves. Most of the light that we are dealing with is anything but sine-shaped. We will come back to this topic in the next chapter.

Exercises

1. The standing waves that we obtain when two waves move against one another are a special case of the interference phenomena that we have just discussed. What is the value of the angle α in this case? Will there be the correct distances between antinodes and nodes?
2. At which angle do two light waves have to intersect (we assume $\lambda = 550$ nm) so that neighboring amplification stripes on a screen have a distance of 2 mm?

4.14 The diffraction of waves

The movement of a wave differs from that of a body in one essential aspect. The body maintains its shape. It is easy to indicate how it moves from a point A to a point B. If we want, we can even indicate a „trajectory“ for each of its points, Fig. 4.42.

Things are different for a wave. A wave does not only change its place but also its shape. There is no tra-

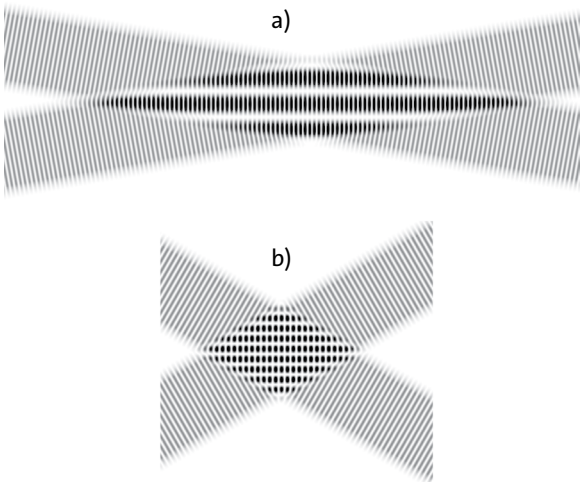

Fig. 4.40 Two sine waves intersect at 20° (a) and at 60° (b). The smaller the angle, the longer the distance between adjacent amplification areas.

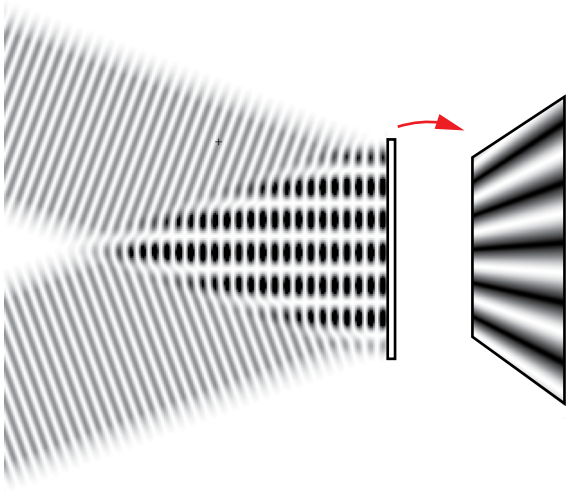

Fig. 4.41 On the screen that we put in the way of the light, we can see bright and dark stripes. (In our Figure perpendicular to the drawing plane)

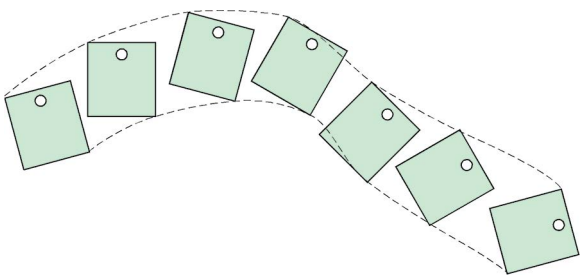

Fig. 4.42 Each point of a moving body describes a specific trajectory.

jectory at first. What comes closest to a trajectory would be a line that always stands perpendicularly on the wave fronts. Let's look at the speaker A and the person B in Fig. 4.43. To the question about the way that the sound takes from the speaker to the person we could answer: a straight way because the straight connection line between A and B stands orthogonally on the wave fronts everywhere.

The example suggests that the wave movement is always linear. Both for a plane wave as well as for a spherical wave, the orthogonal lines are straight lines. This is what we claim when we describe light by means of rays.

Now we look at the situation of person C. Music comes out of the speaker and the person hears this music. But the sound waves do certainly not go from A to C on a straight line because they would have to move through the wall in that case. They move around the wall. Hence, the path of the wave is no longer straight but curved.

We say that the waves are *diffracted*. Therefore, waves can move on curved ways in this sense.

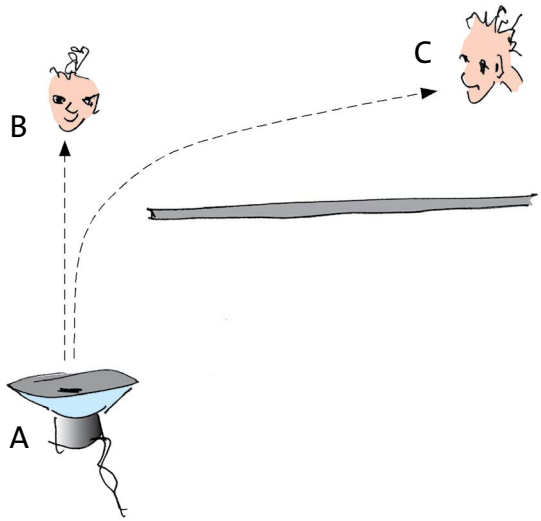

Fig. 4.43 The sound can move on the straight connection line from A to B. From A to C, the sound waves run around an obstacle.

Diffraction

A wave moves from A to C although there is an obstacle on the straight connection line between A and C.

Diffraction is a phenomenon that occurs to all waves. We have already addressed the sound: to hear something, we do not need a straight „line of sight" to the sound source. Although the waves that come from the open sea are held back by a pier, something of the swell can still be felt behind the pier. Radio reception is also possible at a place where there is no line of sight to the sender, albeit to a varying extent depending on the wavelength range.

Things only seem to be different for the light. Light only moves to places that it can reach on a straight path. This is the reason why we can create harsh shadows and the reason behind the existence of *light rays*. But how can this be compatible with the fact that waves are diffracted on obstacles?

We look once again at a plane wave that moves partially against a wall that extends into its path, Fig. 4.44.

The wave is diffracted on the edge of the wall. It also moves into the area behind the wall. But not all of the light that arrives on the left of the edge will be diffracted. The impact of the edge becomes increasingly weaker towards the left. We can memorize as a rule:

The part of a wave whose distance to the obstacle is approximately one wavelength is diffracted.

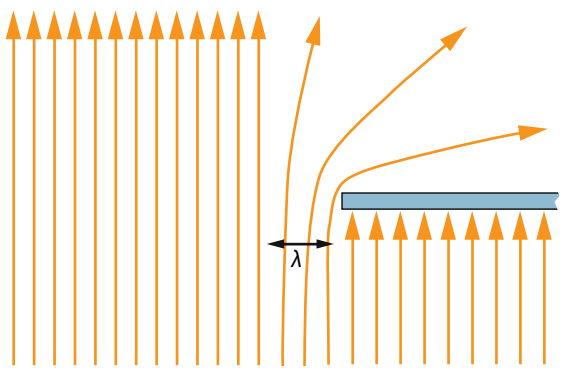

Fig. 4.44 The part of the wave that reaches into the „shadow area" comes from an area whose width is approximately one wavelength.

The longer the wavelength, the more of the incident wave is therefore diffracted into the shadow area.

Now we can see why light waves are diffracted so little compared to sound waves: light has a much shorter wavelength than sound.

Nevertheless, there are situations in which we can clearly see the diffraction of light. In a thin wall, there is an opening with a diameter of only a few light wave lengths. We illuminate the pinhole from one side. On the other side, i.e. behind the wall, instead of the pinhole we can see a bright spot, also when looking from the side at an angle. As the opening is small, not much light can pass it. But practically all the light that passes will be diffracted.

We can therefore also understand the conditions under which light can be regarded as rays: all openings delimiting a light wave have to be large compared to the wavelength. But the width of a light wave is not only constricted by openings (so-called apertures) but also by the edges of mirrors and lenses.

As we know, the law of reflection („angle of incidence equal to angle of reflection") applies for mirrors. But if a mirror is too small, the law will no longer apply. The light will be diffracted on the mirror edges. Accordingly, this is true for refraction.

For the laws of „geometrical optics" to apply, the diameter of apertures, lenses and mirrors has to be large compared to the wavelength.

The laws of geometrical optics can just as well be applied to radio waves and sound waves. We only need to use mirrors and lenses that are sufficiently large.

For a „wall" to reflect an electromagnetic wave, its surface must be electrically conductive. The shorter the wavelength, the smoother the mirror surface needs to be. Irregularities have to be significantly smaller than the wavelength. If the irregularities have a similar size as the wavelength, the wave will be *scattered*, i.e. diffracted in a variety of directions. The „mirror" will no longer be a mirror.

Very remote galaxies are observed inter alia by means of the radio waves they emit. Typical wavelengths are in the range of several meters. Parabolic mirrors with a diameter of up to 100 m are used to bundle these waves. Due to the long wavelength, such mirrors can be made of meshed wire.

Exercise

1. A normal television antenna is mostly set up in a way that there is no line of sight to the television channel. From a parabolic antenna, in turn, there has to be a straight, unobstructed connection to the satellite. Explain the difference. Why do clouds not disturb the reception with the parabolic antenna? Why do mobile phones and wireless landline phones also work behind a wall?

4.15 The elementary portions of sound waves, electromagnetic waves and matter waves

It would actually be much more convenient if this last section were not needed. The world would be clear and simple: light and sound are waves on a carrier that

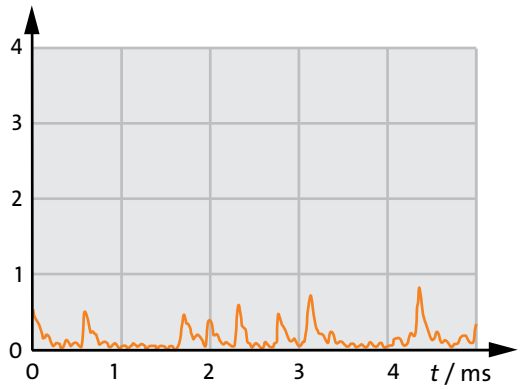

Fig. 4.45 The photomultiplier registers portions of light.

can be imagined as a substance that is distributed completely evenly, i.e. a continuum as we say. Unfortunately, however, natural science has demonstrated time and again that, after having quite understood a topic, new problems, which could no longer be solved with the methods and tools used up to then, emerged. This was also the case for the sound and the light and for other waves. The evidence of them being wave phenomena is unambiguous, but there are still observations that cannot be explained based on the concept of waves. They even appear to be contradictory to the idea of waves.

Actually, light sometimes behaves like a current of small bodies that are flying at irregular distances.

This can be seen in the diagram from Fig. 4.45. It shows the „signal" that a highly sensitive light meter (a *photomultiplier*) provides when it is hit by an extremely weak light beam. As light is a wave, we might expect the meter to provide an even, albeit weak, signal. In case the meter reacts sufficiently fast, we could possibly also expect a sineshaped signal. But neither the first nor the latter happens. Small portions of light are registered in irregular intervals.

We therefore have to get used to the fact that light may exist in different states, i.e. sometimes clearly as a wave and in other cases clearly as a small portion or „particles", but in most cases somewhere in between. And this also applies for the other wave phenomena, for example the sound. But the fact that something behaves sometimes like a wave and sometimes as if was formed of particles does not only apply for the phenomena that we have gotten to know as waves. It also holds true for the entities that usually appear to us as particles: electrons, protons, atoms, etc. Under specific circumstances, these objects also appear to us as waves.

For the two extremes – waves and particles – we have quite simple description methods. But there is also a theory that describes the light – and also the other phenomena – in all states: quantum mechanics. It is necessarily more complicated and less intuitive than the theory of waves and that of particles. Only when dealing with quantum mechanics, you will understand under which circumstances something behaves like a wave and under which circumstances it behaves like a particle. Table 4.2 shows the names of the waves with those of the associated particles.

Here you can already see some rules that will allow us to connect one aspect to another.

In all cases, the energy E of the particles is connected in a simple way to the frequency f of the associated wave:

$$E = h \cdot f$$

Here, h is Planck's constant:

wave	particle
electromagnetic wave	photon
sound wave	phonon
matter wave	electron, ...
gravitational wave	graviton

Table 4.2 Names of waves wave particle and associated particles

$$h = 6.626 \cdot 10^{-34}\ \text{Js}$$

The relationship between the momentum p and the wavelength λ is only a little more complicated:

$$p = \frac{h}{\lambda}$$

We would like to leave it at that for the moment. You will read more about it later in the context of quantum mechanics.

5 INTERFERENCE OF LIGHT AND X-RAYS

Interference effects, i.e. the amplification and attenuation that occur during the superposition of two or more sine waves, are an essential instrument of physical research.

The light originating from excited atoms and molecules when they return to the ground state can be analyzed very accurately by means of interference measurements. Therefore, the most important instrument to explore the structure of atoms and molecules is at our disposal.

The structure of solid substances and large molecules consisting of atoms are examined by means of the interference of X-rays, i.e. with electromagnetic waves with very short wavelengths. In this context, the complicated substances that we come across in molecular biology are particularly interesting.

Another field of application is astrophysics. For example the diameter of remote stars is often measured by means of the interference of light.

In the following, we will therefore look at the interference of light and X-rays. You might think that there is nothing fundamentally new anymore. But there is a problem in connection with these radiations that will cause us some trouble: in the previous sections we have studied only sine waves. In reality, however, we never come across pure sine waves.

5.1 Coherence

Again: to cause interference of light, we need two or more sine waves. And this is our problem because the light that is usually available to us is anything but sine-shaped. For example the light on a foggy day only consists of untidy wave ripples. A corresponding two-dimensional wave would look approximately like that illustrated in Fig. 5.1.

Now we recall a rule that we have learned in connection with oscillations.

> Every function can be expressed as a sum of sine functions.

At first, we look at a wave with straight wave fronts that moves in the x-direction, for example that from Fig. 5.2. We can see that it is not a sine wave. But we

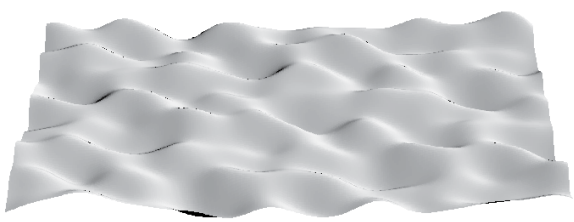

Fig. 5.1 This is how irregular light waves are outdoors on a foggy day.

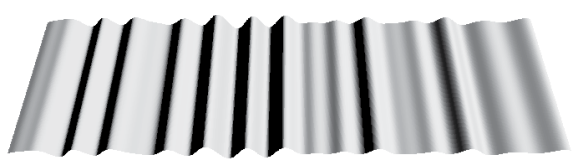

Fig. 5.2 This „plane" wave can be decomposed into plane sine waves.

can conclude from the rule cited above that we may imagine the wave to be composed of sine waves with different wavelengths and amplitudes. In mathematical terms: the function $y_{t0}(x)$, which describes a snapshot of the wave at the instant of time t_0, can be expressed as a sum of sine functions. In fact, the wave from the figure was created by adding up 5 sine terms.

The rule can even be generalized. Also a wave that has no straight wave fronts can be decomposed in sine waves. But in this case, the sine wave components must also have different directions.

> Each wave can be composed of sine waves with different amplitudes, wavelengths and directions.

The wave ripples from Fig. 5.1 were created through the addition of 10 sine waves with different directions, wavelengths and amplitudes.

We would like to obtain a few rules to assemble a wave from sine waves. We examine the problem by adding up 5 sine functions. The rules that we find, however, will also be valid if many more sine waves are added up.

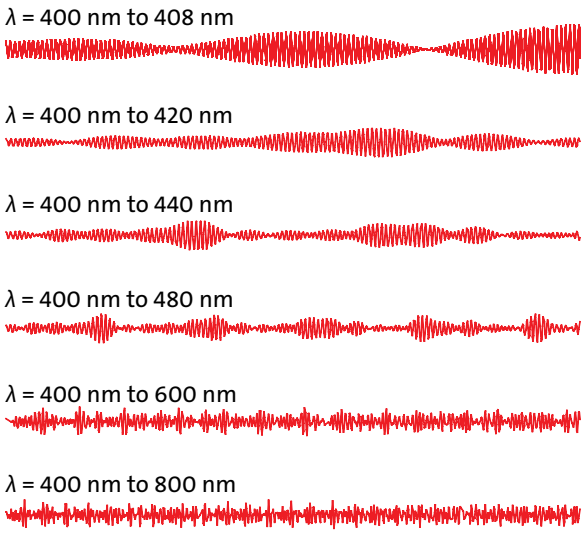

Fig. 5.3 Each of the waves is composed of 5 sine waves. The λ range of the wave components is very small at the top, i.e. 2 % of 400 nm, and increases downwards up to 100%.

1. Sine waves with different wavelengths

We start with the addition of 5 waves of the same direction and illustrate the result in one dimension, i.e. the wave quantity y over the position x.

At first, we take waves whose wavelengths only differ by a small amount. All five wavelengths are in a narrow interval $\Delta\lambda$. They differ by a maximum of 8 nm. The result is displayed at the top of Fig. 5.3.

Then, we choose an increasingly larger $\Delta\lambda$ range: 20 nm, 40 nm, 80 nm, 200 nm and finally 400 nm. We can see: the larger the λ range, the less the resulting wave resembles a sine wave. While we can still identify larger continuous sine-like sections at the top, the wave only consists of an untidy up-and-down at the very bottom. When the wavelengths of the wave components are only slightly different from each other, we obtain long pieces that roughly look like sections of sine waves. The length of these pieces is called *coherence length* (from the Latin word cohaerere: to be connected). For the upper wave, it amounts to approximately 20 to 30 wavelengths. The wave behaves like a continuous sine wave over approximately 25 wavelengths; then, it goes out of sync. For the fourth wave from the top, the coherence length is approximately 3 wavelengths; in case of the penultimate wave, we might detect – optimistically speaking – one sine period, and in the case of the wave at the extreme bottom, the sine character has disappeared completely. The coherence length λ_{coh} in wavelength units can be obtained as a quotient of λ and the interval $\Delta\lambda$:

$$\frac{\ell_{coh}}{\lambda} = \frac{\lambda}{\Delta\lambda}.$$

For λ, we insert the mean wavelength of the wavelength interval. It is not worth calculating the coherence length with a very high accuracy. Let's use the 3rd wave from the top in Fig. 5.3 as an example. Here, we have $\lambda \approx 420$ nm and $\Delta\lambda = 40$ nm. The coherence length in wavelength units turns out to be

$$\frac{\ell_{coh}}{\lambda} = \frac{\lambda}{\Delta\lambda} = \frac{420\ \text{nm}}{40\ \text{nm}} = 10\,.$$

We thus have the following rule for assembling sine waves:

> Sine waves from a small wavelength range:
> long coherence length of the resulting wave
> Sine waves from a large wavelength range:
> short coherence length of the resulting wave

2. Sine waves with different directions

Now we compose a wave of sine waves with different directions but with the same wavelength. Again, we take 5 components and again, they are very similar to one another at the start: the angle $\Delta\alpha$, by which their

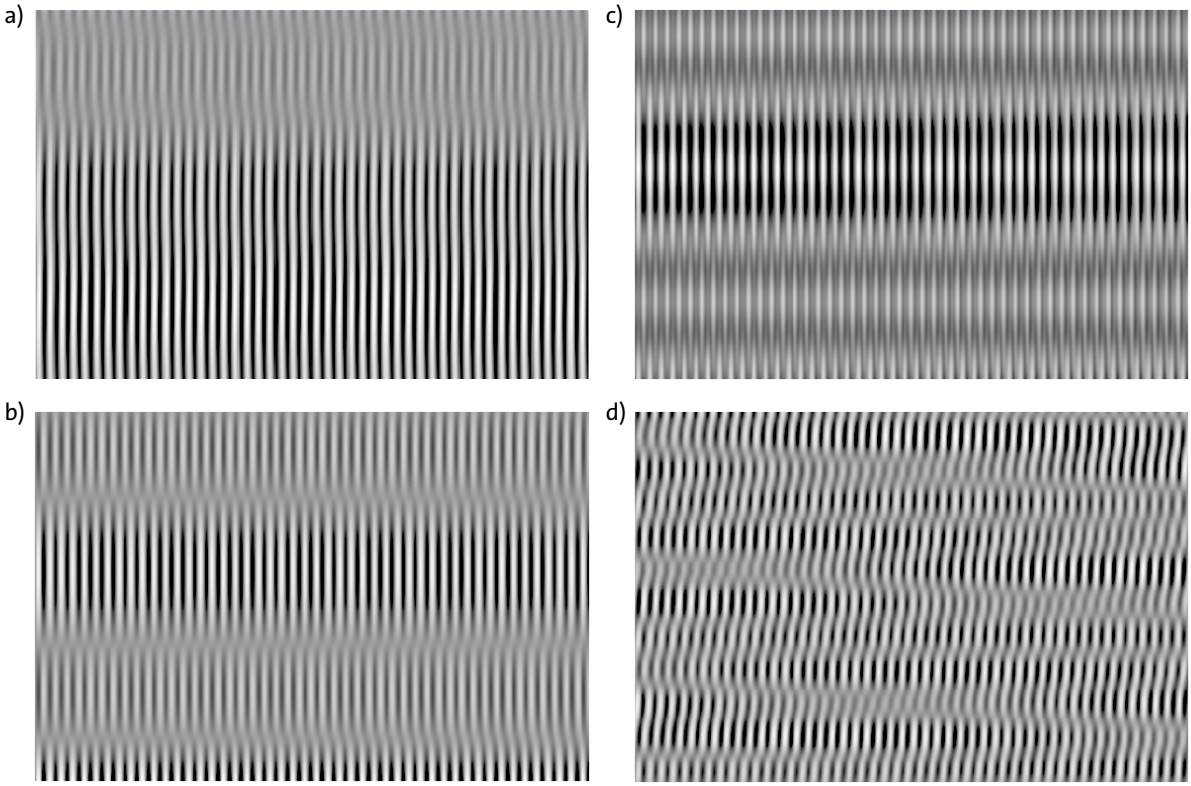

Fig. 5.4 Assembling 5 sine waves of the same wavelength but different directions. The angular range of the waves is 2° (a), 5° (b), 10° (c) and 20° (d). The smaller this range, the larger the width of the connected areas, i.e. the coherence width.

directions differ as a maximum, is at first 2°, Fig. 5.4a. Then, we choose $\Delta\alpha = 5°$, 10° and finally 20°.

Again, we can observe connected areas that look like sections of sine waves. This time, however, they are not limited in their length but in the width. This width is the coherence width. In the first case it is largest. With increasing $\Delta\alpha$, the coherence width decreases more and more.

> Sine waves from a small angular range:
> long coherence with of the resulting wave
> Sine waves from a large angular range:
> small coherence width of the resulting wave

3. Sine waves with different wavelengths and directions

Now we compose a wave from sine waves with different wavelengths and different directions. The wave from Fig. 5.5 consists of 6 different sine waves, of a relatively small wavelength range and a small angular range.

We can now identify coherence regions of a certain length and a certain width.

Let's describe these findings once again and in a different way.

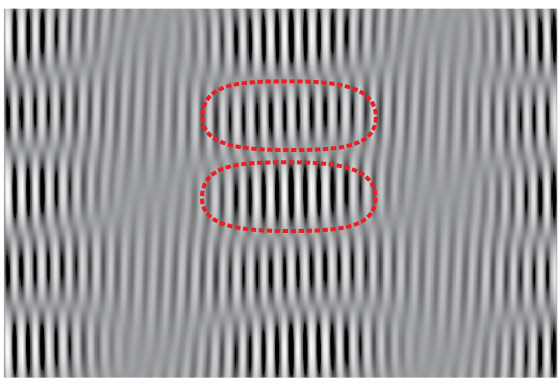

Fig. 5.5 Assembling 6 sine waves of different wavelengths and different directions. Two coherence ranges are marked by dashed lines.

A wave generally consists of a mishmash of sine waves. There are two contributions to this mishmash:

- a mishmash of wavelengths;
- a mishmash of directions.

Light of a variety of compositions can be found in nature and technology. We will look at a few examples.

1. Fog

You are surrounded by fog. Light comes from all directions, and all wavelengths are represented, i.e. from approximately 400 nm to 800 nm. Hence, the light is completely disordered, both regarding the directions as well as the wavelengths. A snapshot would look approximately like Fig. 5.1. No coherence ranges whatsoever can be identified. We say that this light is completely *incoherent*.

2. A remote light bulb

The light comes from a very small angular range, i.e. practically from a single direction, but there are sine waves with a variety of wavelengths. The light looks approximately as illustrated in Fig. 5.2: a very large coherence width (the width is larger than the displayed wave piece), but no coherence length exists. Hence, it is highly ordered in terms of direction but a mishmash as far as the wavelength is concerned. Light that is completely incoherent with regard to the wavelength is formed in *thermal* light sources: bodies that are made glow by means of increasing their temperature to a sufficiently high level. Examples are the Sun, all other stars, light bulbs and other glowing bodies.

3. Orange street lamp and fog

Again, let's assume there is fog. You are standing close to a street lamp that creates orange light. We assume that it consists of sine waves with a single wavelength. (Strictly speaking, this is slightly exaggerated. Although the orange light of street lamps has a narrower spectrum than the light of a light bulb, it has more than just a single wavelength.) The sine waves of our assumed light consequently have a single wavelength but very different directions. With regard to the wavelength, the light is highly ordered, but chaotic in terms of its direction.

4. Laser

The light has a uniform direction and a uniform wavelength. It is a single sine wave. It is ordered with regard to the wavelength and the direction, i.e. it is *coherent*. Strictly speaking, the coherence length is not unlimited in case of the laser either. Even laser light goes out of sync from time to time. A typical value for the coherence length of laser light is 1 m. This corresponds to no less than 100,000 wavelengths.

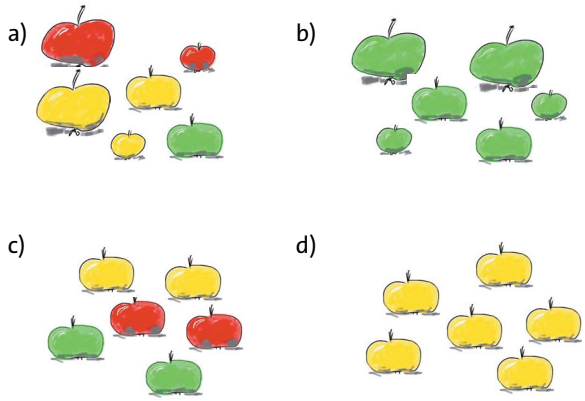

Fig. 5.6 Apples: (a) different colors, different sizes; (b) one color, different sizes; (c) one size, different colors; (d) one color, one size

As a conclusion, consider the following parable.

We compare the light with apples. Apples can also be a mishmash in two respects, Fig. 5.6. First, they can have different colors, and second, they can have different sizes. A box with apples of different colors and sizes corresponds to the completely incoherent light. A box with apples of a single color but different sizes would correspond for example to the light from the remote light bulb, i.e. the light with the mishmash of wavelengths. A box with apples of a uniform size but different colors would correspond to the light of the orange street lamp in the fog. A box with apples of a uniform size and color (like the ones at the supermarket) would correspond to the completely coherent laser light.

Exercises

1. The spectrum of a red light-emitting diode ranges from approximately 640 to 650 nm. What is the coherence length in wavelength units?
2. The yellow light of a sodium flame has a wavelength of approximately 590 nm and comes from a wavelength range of $\Delta\lambda = 0.6$ nm. What is the coherence length in mm and in wavelength units?
3. Radio channels are assigned frequency intervals. In the sender, the signals to be broadcast (music, speech, etc.) that only contain the sound frequencies are transformed into a mix of waves from the range of the assigned frequencies (they are „encoded"). The signals of SWR2 in Stuttgart are approximately in the range from 105.65 and 105.75 MHz. What is the coherence length of the radio waves?

5.2 How to produce coherent light

To obtain interference, pure sine waves are needed. Hence, coherent light. But what can be done if only a completely incoherent light is available? How can a pure sine wave be obtained from such light?

There is only a single method, and it is the same as that described with the apples. We imagine to have a large box with apples of diverse sizes and colors, i.e. a mingle-mangle in two respects. However, only the big red ones can be sold. What can we do? It is very simple: we take out the big red ones and put all the other ones aside, maybe to make apple juice.

Things are similar for the light. If we have a mix of sine waves but only need one type, we will have to filter out the waves that we need. All remaining ones are useless. We cannot transform incoherent into coherent light, just as we cannot transform red into green apples or big into small ones.

So how can we then blind or filter a single sine wave out of a sine wave mix, like that supplied by a normal lamp?

Let's first extract light with a single direction (or better: light from a small angular range) from light with many different directions. Fig. 5.7 shows a possibility: putting two small pinholes in a row. Only waves that have the direction of the connection line between the two holes can pass the second hole. A second method: we move far away from the light source. In the environment of any observation point P, we have a nearly plane light wave that moves in the direction of the straight line that connects the light source to P. The light waves that come from a star are perfect plane waves in the small ranges that are interesting to us.

Let's assume that we have sorted out the undesired directions; our wave has now a unique direction. But there is still the chaos of wavelengths. There are many methods to filter out a sine wave of a single wavelength, or more precisely: sine waves with a small wavelength interval. The cheapest one is (as the word „filter out" already indicates) a filter: a glass plate that absorbs or reflects the greatest part of the light and that only lets light from a small wavelength range pass.

The analogy with the apples can once again be helpful. If only the big red apples could be sold, it would be better to plant only apple trees whose apples are big and red from the outset.

The same applies for the light: if we need coherent light, it will be best to use a light source that only supplies coherent light from the start: a laser. In earlier times when there was no laser, people had to get by with the methods that we described before.

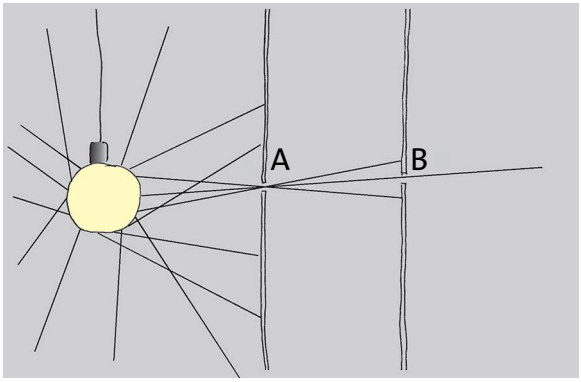

Fig. 5.7 At the point A, the light consists of sine waves of diverse directions. At B, it is a nearly plane wave.

5.3 Even laser light is not sufficient

Again, we explain the problem by means of two plane waves that move towards each other at an acute angle and that move through one another, Fig. 4.33. We discussed this experiment theoretically in the previous chapter, but we abstained from trying it out – because it would not have worked. But how should it have been done then? We could assume that nothing is easier than that. We take two lasers, i.e. sources of coherent light, direct the beams onto each other at an acute angle and set up a screen in the intersection area. Bright and dark stripes should be visible on the screen. In fact, however, we can only see a medium-light spot. Why?

This is because lasers do not do what we expect from them. Although every laser makes a wave that looks like a sine wave over many wavelengths, it still goes out of sync after a certain time. And a bit later again...and this happens to both lasers, regardless of each other. We could imagine it best as the laser making a jump in the phase at a random time as shown by the „sine function" of Fig. 5.8.

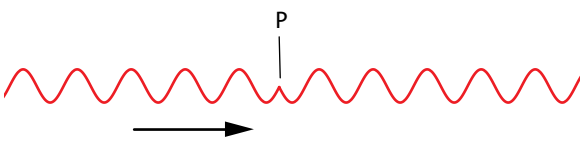

Fig. 5.8 A wave goes out of sync at P. It makes a „phase jump".

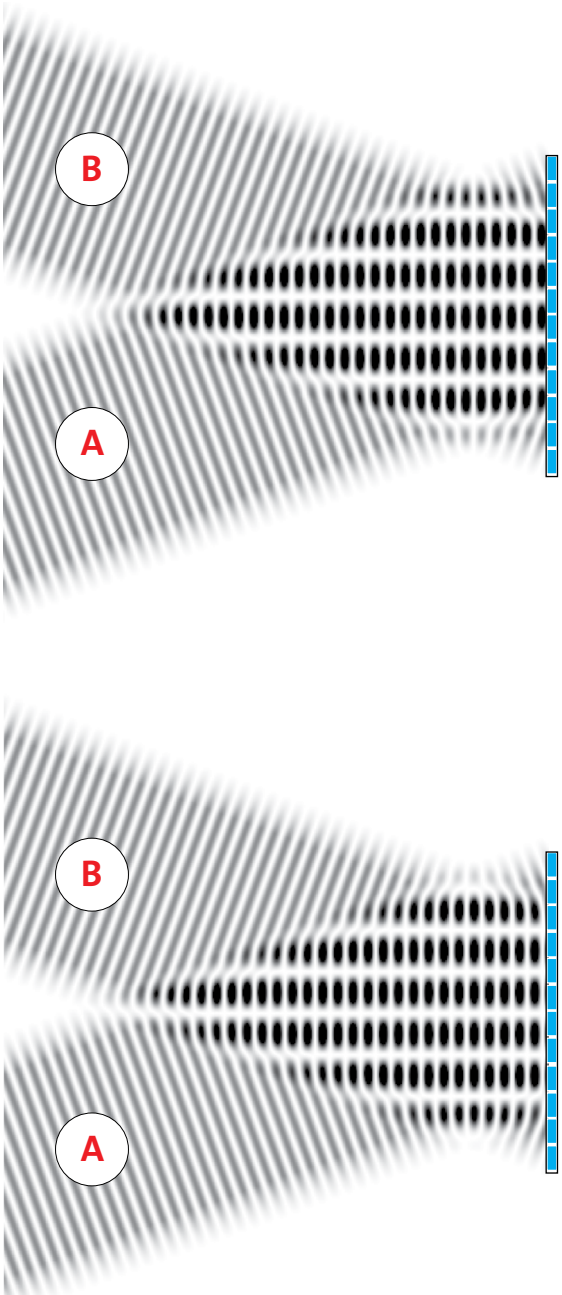

Fig. 5.9 Two sine waves interfere. On the screen (right side) we can see a sequence of bright and dark stripes. In the lower image wave A was displaced by half a wavelength in relation to B. As a consequence the stripe pattern was offset. Where there are bright spots on the screen in the upper picture, it is dark in the lower picture and vice versa.

What is the effect of this on the interference picture on the shield? What will happen to the interference pattern if for example one of the waves is displaced by half a wavelength? Fig. 5.9 shows two interference processes. The difference is that the wave A in the lower picture is displaced by $\lambda/2$ in relation to wave A in the upper picture. The result: also the interference pattern was offset. Where there is extinction in the upper picture, there is amplification in the lower one and vice versa.

Hence, each time a wave goes out of sync, the stripe pattern on the screen jumps to one or the other side. The process of going out of sync does not have to happen abruptly. It can also be a bit more unhurried, because in that case, also the interference stripes move steadily back and forth. This is exactly what happens in reality, even when we use very good lasers. The back-and-forth jumping movement of the interference stripes is so fast that we can only perceive the mean value with the naked eye: we observe an even, textureless brightness.

After these long deliberations, we are finally able to set up an interference experiment that actually works. As we cannot find two light sources that stay in sync for a long time, we have to use two sine waves from the same source. If one of them will go out of sync, the other one will go out of sync at the same time and the interference pattern will not be offset in the process. There are many possibilities to make two waves out of one; an example is shown in Fig. 5.10. The light wave of a laser falls onto a „mirror“ that reflects half of the light and lets the other half pass. The waves A and B do not only have the same coherence length, but each deviation from a sine wave in A corresponds to an identical deviation in B. Hence, wave A stays in sync in relation to wave B. When A and B intersect, an interference pattern that does not jump back and forth is formed.

There is only one further aspect to consider: the paths that A and B travel from the semi-transparent mirror to the region of interference must not be very different. If wave B has a longer way than A, it will hit sections of A that are still part of the preceding coherence range. The path difference between A and B has to be in any case smaller than the coherence length. But this condition is usually easy to fulfill.

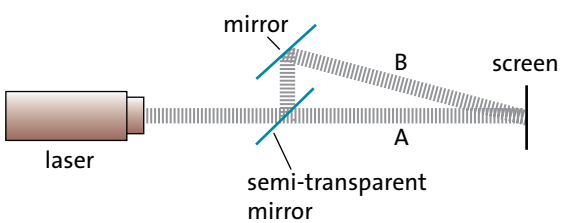

Fig. 5.10 Waves A and B interfere with each other. A stripe pattern can be seen on the shield.

Even if we were not to get any further benefit from the interference of light, the experiment described before is of particular importance. It proves that light is a wave. The first interference experiment was made successfully by the physicist Thomas Young in 1801. Up to then it had been assumed that light was a wave phenomenon, but there had not yet been a clear proof.

Exercises

1. (a) Let's assume that the light of a laser has a wavelength of 633 nm and a coherence length of 15 cm. How many wavelengths is the equivalent of the coherence length? How fast does the stripe pattern flicker when we create an interference picture with two such lasers? (b) What would have to be the coherence length in order to make a stripe pattern stay in the same place for 1 second?
2. Interference experiment with sound waves: we create two sine waves with two speakers and examine the wave by means of a microphone. Do we have the same problems as with light in this case? If no, why not? If yes: how can they be bypassed?

5.4 Diffraction by pinholes and slits

Now we get to know an even simpler method to create interference patterns. We have seen that we need two sine waves that either never go out of sync or, in case it still happens, it has to happen simultaneously to both.

We let a sine wave move against a pinhole. The pinhole should be smaller than the wavelength. The wave is diffracted and moves away in all directions behind the pinhole, Fig. 5.11. We have a sort of spherical wave. It differs from a real spherical wave in the dependence of the amplitude on the direction. In a „forward direction", the amplitude is large, and it decreases with a growing angle against this direction.

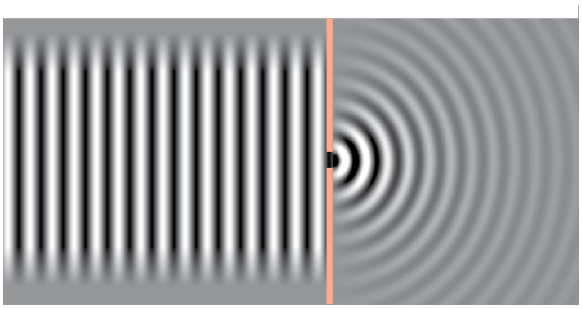

Fig. 5.11 The wave fronts are circular behind the pinhole (and spherical for a three-dimensional wave).

Now we let the incoming wave move against two pinholes. Now, two spherical waves are formed behind the obstacle, and these waves interfere with one another. As both of them originate from the same incoming wave, they will always stay in sync.

The two „spherical waves" and their superposition are displayed as a snapshot on the left in Fig. 5.12. The image on the right shows the time average value of the

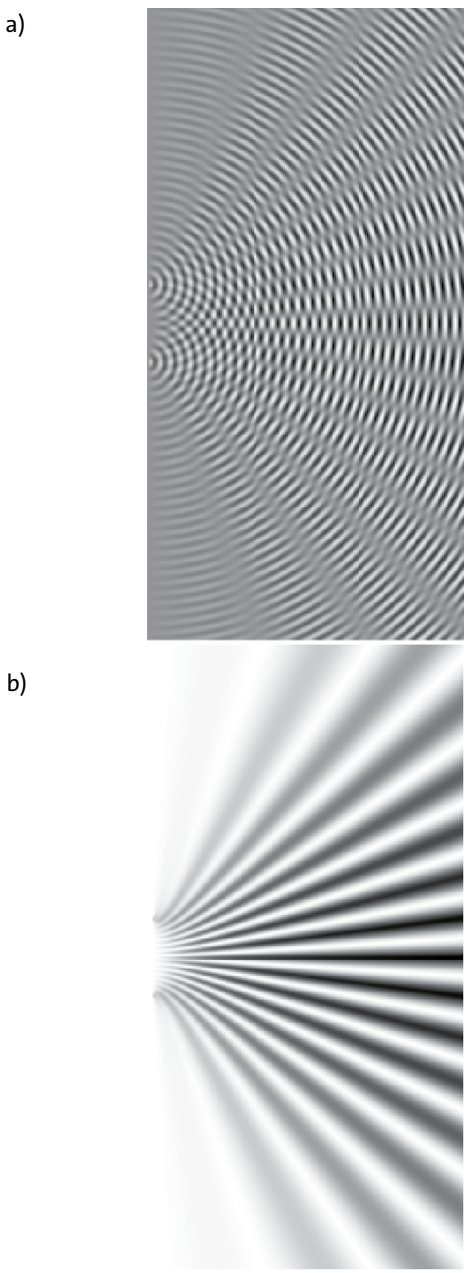

Fig. 5.12 a) Snapshot of the spherical waves that move out of two pinholes. b) Mean value of the square of the wave function.

square of the wave function, i.e. the intensity. Black stands for a high intensity, white for a low one. There is amplification in the black areas and extinction in the white areas. Hence, we have obtained an interference pattern in a very simple way. If we put a screen in the way of the light, we will see bright and dark spots in an alternating way on it.

If we take two slits – in Fig. 5.12, they would be perpendicular to the drawing plane – instead of two pinholes, we will obtain bright and dark stripes as an interference pattern on the screen, similar to those that we created earlier by means of the mirrors.

We would like to analyze in greater detail how these stripes originate and calculate their distance.

A plane sine wave falls from the left onto the double slit in Fig. 5.13. Two waves whose wave fronts are circular in a cross-sectional view originate from the two slits. We only want to see what happens in the plane of the screen: will the waves in a given point P amplify or extinguish themselves?

We connect each of the two slits with P through a straight line. The straight lines are the radii of the circular waves. We can see that the path r_1 of the lower wave is longer than the path r_2 of the upper one. As the waves are in sync at slit 1 and slit 2, they are no longer in sync on the screen. The farther P is away from the center of the screen, the greater will be this path difference. Every time the path difference is an integer multiple of λ, the two waves amplify. When it is $\lambda/2$, $3\lambda/2$, etc., they will extinguish themselves:

Path difference	results in
$0, \lambda, 2\lambda, 3\lambda, 4\lambda, \ldots$	amplification
$\lambda/2, 3\lambda/2, 5\lambda/2, 7\lambda/2, \ldots$	extinction

This can be written in a shorter way:

Path difference	results in
$k \cdot \lambda$	amplification
$(k + 1/2) \cdot \lambda$	extinction

Here, k stands at first for the integers 0, 1, 2 etc. But we can also insert the negative integers for k. A negative path difference means that the path of the upper wave is longer than that of the lower wave.

At the point on the shield where the path difference is 0, there is the *intensity maximum of zeroth order*; the neighboring points of maximum intensity are the two maxima of first order, which are followed by the maxima of second order, and so forth.

We would like to express the path difference by means of the angles that form the two radii r_1 and r_2 with the line that is perpendicular to the slit plane. This appears complicated at first, but it is not. The distance between the slit plane and the screen is usually so big compared to the distance between the slits d that the two radii are practically parallel. Hence, Fig. 5.13 is actually a poor illustration of this matter. When we illustrate the experiment to scale, we have to draw the distance between the slits so small that nothing can be recognized anymore, Fig. 5.14a. Fig. 5.14b therefore only shows the vicinity of the slits in a strongly enlarged way.

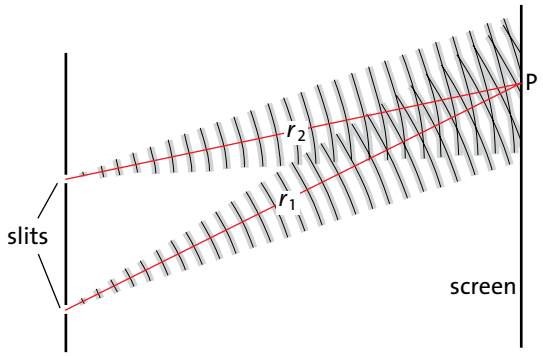

Fig. 5.13 The path r_2 from the upper slit to the point P is shorter than the path r_1 from the lower slit to P. If the path difference is an integer multiple of λ, the two waves will amplify.

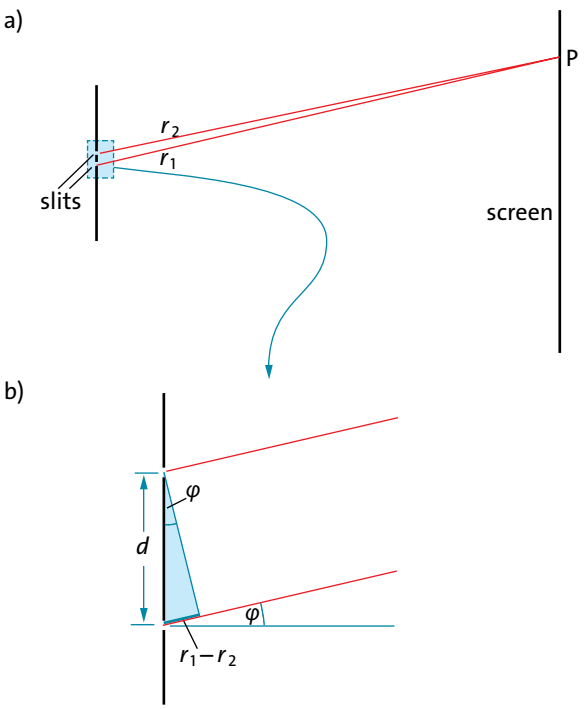

Fig. 5.14 (a) The distance between the double slit and the observation shield is so long that the radii r_1 and r_2 are practically parallel. (b) Enlarged view of the vicinity of the double slit

The angle of the two radii r_1 and r_2 against the line perpendicular to the slits shall be called φ. We can read from the Figure:

$$\sin\varphi = \frac{r_1 - r_2}{d},$$

or

$$r_1 - r_2 = d \cdot \sin\varphi.$$

We can therefore reformulate our rule:

$d \cdot \sin\varphi = k \cdot \lambda$ amplification
$d \cdot \sin\varphi = (k + 1/2) \cdot \lambda$ extinction
with $k = \ldots -2, -1, 0, 1, 2, \ldots$

To each angle φ corresponds a particular point on the screen. When we denominate the distance from the center of the screen with a and the distance between the slit plane and the plane of the screen with l, Fig. 5.15, we obtain:

$$\tan\varphi = \frac{a}{l}. \qquad (5.1)$$

We would like to express our rule for amplification and extinction with a instead of the angle φ. Here, we can once again take advantage of the very long distance between the slits and the shield. This is why a is also very small in relation to l. The angle φ is consequently very small, and the value of the sine function for small angles is approximately equal to the value of the tangent function:

$$\tan\varphi \approx \sin\varphi. \qquad (5.2)$$

With (5.1) and (5.2), our rules for amplification and extinction become:

$$a = \frac{l}{d} \cdot k \cdot \lambda \qquad \text{amplification}$$

$$a = \frac{l}{d} \cdot \left(k + \frac{1}{2}\right) \cdot \lambda \qquad \text{extinction}$$

d = distance between the slits
a = distance from the center of the interference pattern
l = distance slit plane - screen

Here, we have only determined the points where the amplification of the waves reaches a maximum and where they are fully extinguished. In between, the intensity changes steadily from a high value to zero. Fig. 5.16 shows the intensity as a function of the position a on the screen.

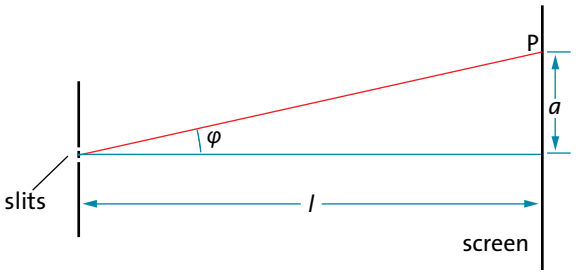

Fig. 5.15 On the relationship between l, a and φ

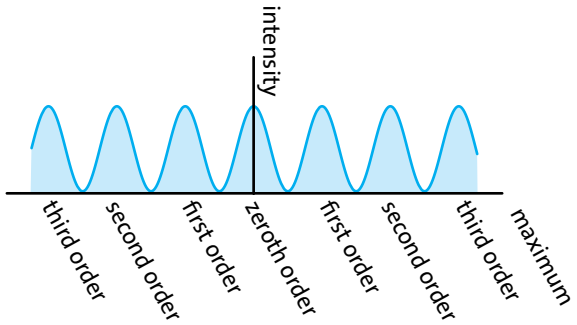

Fig. 5.16 Intensity of the light as a function of the position on the observation screen

Exercises

1. (a) Test with your calculator how well is the approximation $\tan\varphi \approx \sin\varphi$. What is the discrepancy in percent for angles of 1°, 5° and 10°? (b) Show that the approximation follows from the geometrical definition of the sine and the tangent of an angle. (The sine of an angle is the length of the opposite side divided by…, etc.).
2. A plane sine wave of light with λ = 520 nm runs against double slit, with a slit distance of 0.2 mm. What is the distance between the interference stripes, if the screen is located at 1.2 m from the slits.
3. Interference stripes with a distance of 2 cm between them are observed. The screen is at a distance of 8 m from the double slit. The distance between the slits is 0.2 mm. What is the wavelength of the light?

5.5 Diffraction grating – the grating spectrometer

Diffraction by a slit only becomes really interesting when we use not only two but many more slits, e.g. a thousand or ten thousand slits. What will change in that case?

It is not hard to predict where on the screen the interference maxima will appear. Fig. 5.17 shows a section of such a grating. If the length of the path between any slit and the screen differs from that between the neighboring slit and the screen by one wavelength (or by an integer multiple of a wavelength), an intensity maximum will appear on the screen. All waves are in phase at the respective point of the screen.

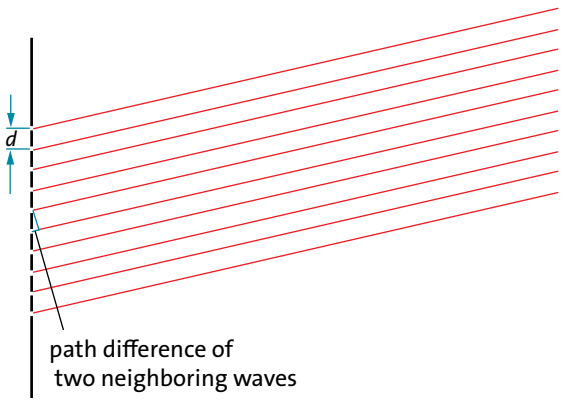

Fig. 5.17 Many parallel slits form a grating. If the path difference between the slit and the observation shield for waves of neighboring slits is equal to $k \cdot \lambda$, the waves will amplify at the shield.

Things are more complicated for the regions between these maxima. If the path difference of the waves of two neighboring slits is $\lambda/2$, respectively two neighboring waves are extinguish. We already know this phenomenon. But there are many other path differences for which the extinction occurs. If the path difference between two neighboring waves is $\lambda/4$, each wave will extinguish itself with its next but one neighbor because the path difference is again $\lambda/2$ between these two waves. If the path difference between neighboring waves is $\lambda/6$, each wave will extinguish itself with its next but two neighbor, etc.. In between, there are path differences for which no complete extinction occurs, but the intensity remains much smaller everywhere than in the maxima. The exact examination is a bit painstaking. The result is shown in Fig. 5.18a for a grating with 10, and in Fig. 5.18b for a grating with 50 slits: the intensity on the screen as a function of the distance from the center.

At the points that correspond to a path difference of λ, there are sharp „peaks“ and in between, the intensity is very low or zero. The more slits there are in the grating, the more waves interfere and the sharper are the peaks and the lower is the intensity between the peaks. In a grating with several thousand slits, practically nothing will be left between the amplification peaks.

This is an important result because it means that a grating can be very helpful: it can be used to decompose a light wave into its sine components. How does that work?

We do not send a sine wave onto the grating, but for example a wave that is composed of two sine waves with different wavelengths. If one of the waves were alone, we would obtain the interference picture of Fig. 5.19a. If the other one were alone, we would obtain the picture from Fig. 5.19b. The composed wave provides the sum of the two intensities, Fig. 5.19c. Hence, we can not only determine the wavelength of *one* sine wave from the interference picture, but also of several sine waves; we can determine the sine components of a complex wave.

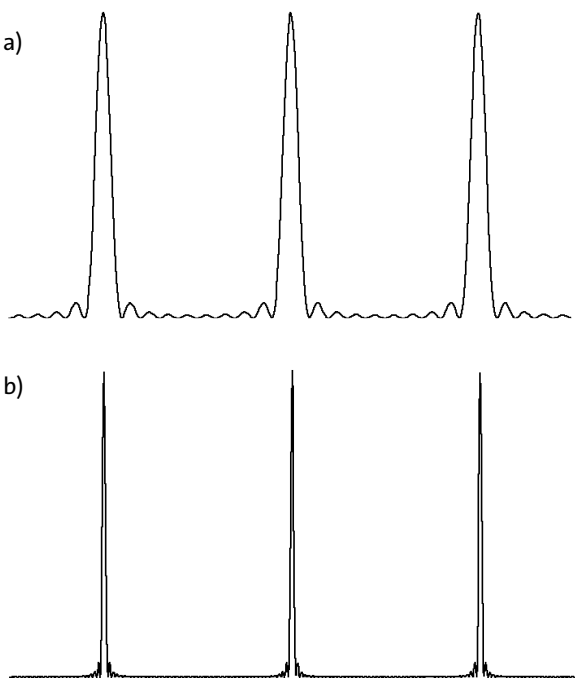

Fig. 5.18 Intensity on the screen for a grating with (a) 10 slits, (b) 50 slits

The grating creates a spectrum on the screen.

If the light contains waves of very different wavelengths, however, the maxima of a sine wave of one order will superpose with those of another sine wave of the next higher order. A grating is therefore only suitable for the analysis of light from a wavelength

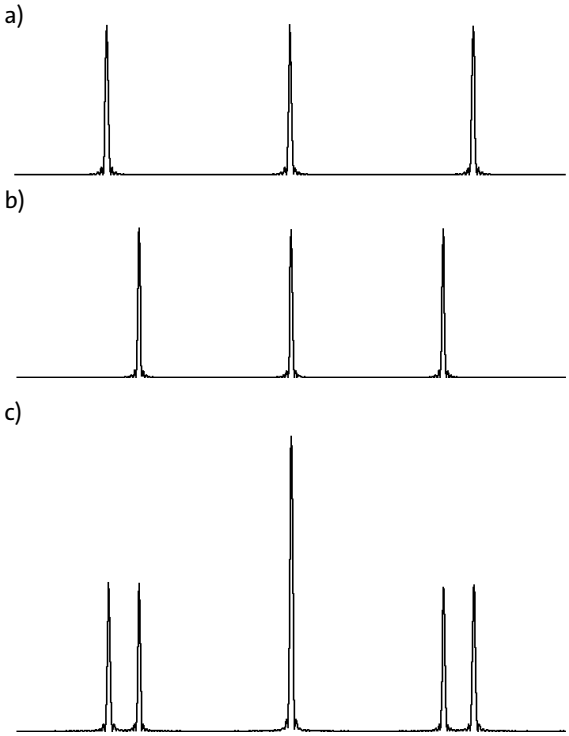

Fig. 5.19 Grating interference picture, created with light of the wavelengths: (a) λ_1, (b) λ_2, (c) λ_1 and λ_2. The maximum of zeroth order is located at the center. The positions of the maxima of first order depend on the wavelength

range that is not too large. Its advantage over the decomposition of light with a prism, is the high *resolution*: we can still measure very fine structures in a spectrum. The device that takes advantage of the interference caused by a grating is called *grating spectrometer*. It is an essential instrument to examine the structure of the atoms and molecules. The major advances of atomic physics, molecular physics, solid state physics and quantum mechanics in the past century are largely based on measurements with the grating spectrometer.

The grating in a technical spectrometer has a diameter of several centimeters and it typically has 1800 lines per millimeter. This means that several tens of thousands of sine waves are brought to interference.

The „gratings" that we have examined so far consisted of slit-shaped openings in an otherwise nontransparent platelet. The grating can also be set up in a way that not the transmitted but the reflected light is brought to interference. This means using a platelet on which thin reflecting stripes are mounted. It absorbs the light between the stripes. Also these reflecting stripes can be regarded as light sources that emit waves with round wave fronts. The same interference picture as for a slit grating is formed, not behind the grating but in front of it, i.e. on the side the light comes from. In order to prevent the light from being reflected back to the light source, one ensures that it falls onto the reflection grating at a slight angle. The zeroth intensity maximum will then be precisely in the direction in which the light would be reflected by a continuous mirror. All CDs are simple versions of such a grating. There are continuous reflecting areas between the data tracks. You have certainly noticed already that this „grating" decomposes the light into its spectral components.

Exercises

1. The light reflected from a CD look quite confusing at first sight. But this is not surprising because the light that hits the CD usually comes from a variety of directions. Try to establish conditions under which you can clearly identify the spectra that correspond to the interference maxima of first and second order.
2. (a) The light of a laser pointer is directed onto a diffraction grating with 300 slits per millimeter. A line pattern is formed at a distance of two meters with the line distance being 32 cm. What is the wavelength of the light? (b) The grating is replaced by another grating with an unknown distance between the slits. The interference lines now have a distance of 48 cm from each other. What are the distances between the slits?

5.6 Two- and three-dimensional gratings

We have assumed up to now that the distance between the slits of the grating is known and that the wavelength of the light is unknown. We were able to calculate the light wavelength from the interference pattern. But also the opposite case can occur: the light wavelength is known and we look for the slit distance. This is the situation when dealing with a different application of interference: the exploration of the structure of crystalline substances by means of X-rays. Also this method is based on the interference of waves that are diffracted on a „grating". However, this grating does not consist of a series of slits, i.e. it is not a monodimensional grating but it extends into the three dimensions of space. To understand the particularities that occur, we use a gradual approach. After the monodimensional grating that we already know, we will at first look at two-dimensional gratings and eventually three-dimensional ones.

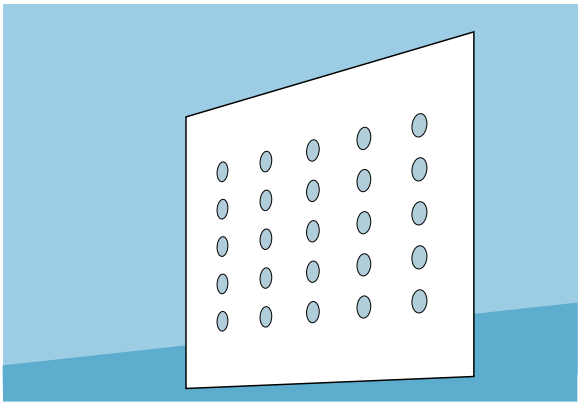

Fig. 5.20 Two-dimensional grating

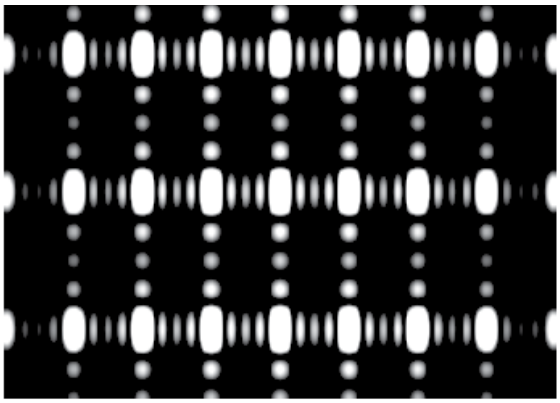

Fig. 5.21 Interference picture of the grating from Fig. 5.20

For the normal, monodimensional grating, we obtain as an interference pattern stripes on the screen that are parallel to the slits of the grating. A two-dimensional grating consists of a platelet with intersecting rows of pinholes or of small mirrors on an absorbing background, Fig. 5.20.

As an interference pattern on the screen, we will not only obtain a structure in the horizontal but also in the vertical direction, Fig. 5.21. As the grating from Fig. 20 only has a width and a height of five points each, the maxima are not yet very clear and there are still points between them at which no complete extinction occurs.

Here we can clearly observe what the formula tells us about the position of the maxima:

$$a = \frac{l}{d} \cdot k \cdot \lambda$$

The shorter the distance between the slits, the greater the distance between the interference maxima. In Fig. 5.20, the distance of the diffracting points is half as long in the vertical direction as in the horizontal direction. Therefore, the distance of the intensity maxima is twice as long in the vertical direction as in the horizontal direction.

Now we pass from the two- to the three-dimensional grating. The points from which spherical waves originate are located on a three-dimensional grid. Fig. 5.22 shows a section of such a grating. During the transition from the monodimensional stripe grating to the two-dimensional point grating, the condition for observing amplification has become more stringent. Instead of lines, only points were visible on the screen. When passing from the two- to the three-dimensional grating, the condition becomes even stricter. It appears as if no amplification could be expected anymore at all.

We let the sine wave obliquely onto the grating and imagine at first that there would only be the first plane

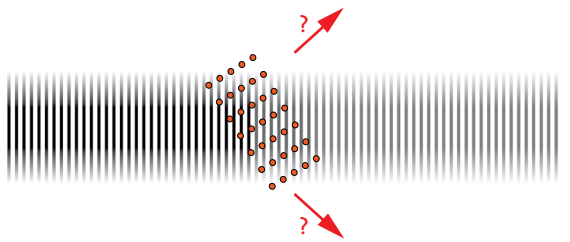

Fig. 5.22 Cross-section through a three-dimensional point grating. A sine wave comes from the left. The biggest part of the wave moves straight ahead through the grating. In which directions will the diffracted waves amplify?

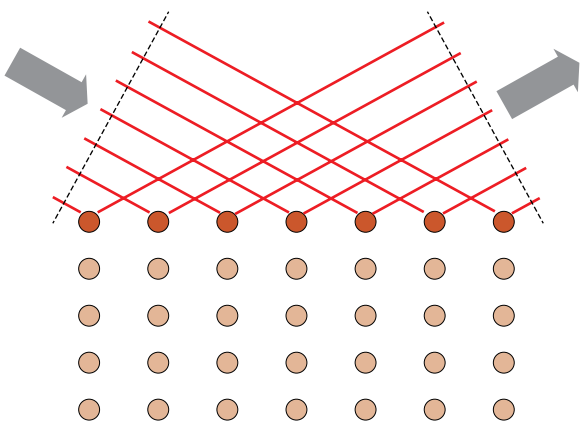

Fig. 5.23 Diffraction at the upper plane. The indicated paths have all the same length because the angle of reflection is equal to the angle of incidence. Without the underlying planes, we would have amplification in the respective direction: the zeroth intensity maximum.

of diffraction centers, Fig. 5.23. The zeroth intensity maximum is located in the direction that can be deduced from the law of reflection (angle of incidence =

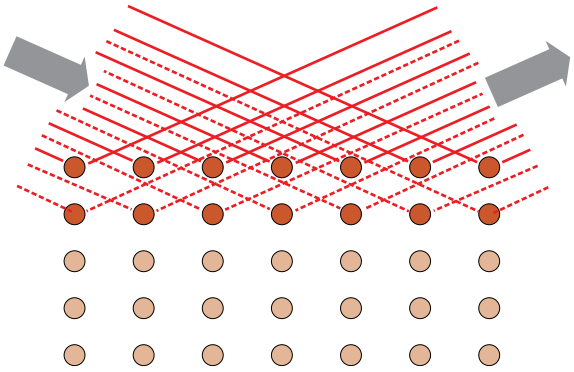

Fig. 5.24 The undashed and the dashed paths have different lengths. There is a phase difference between the corresponding waves.

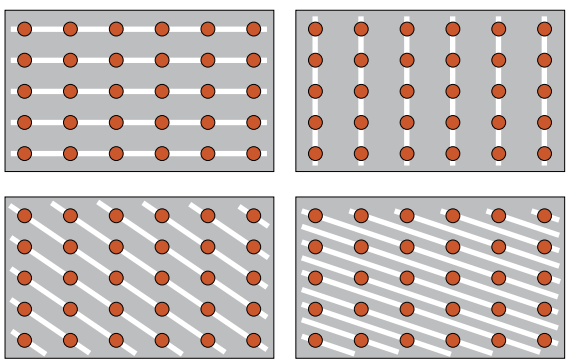

Fig. 5.25 Each set of planes creates intensity maxima, provided that the grating is at a suitable angle to the incident wave.

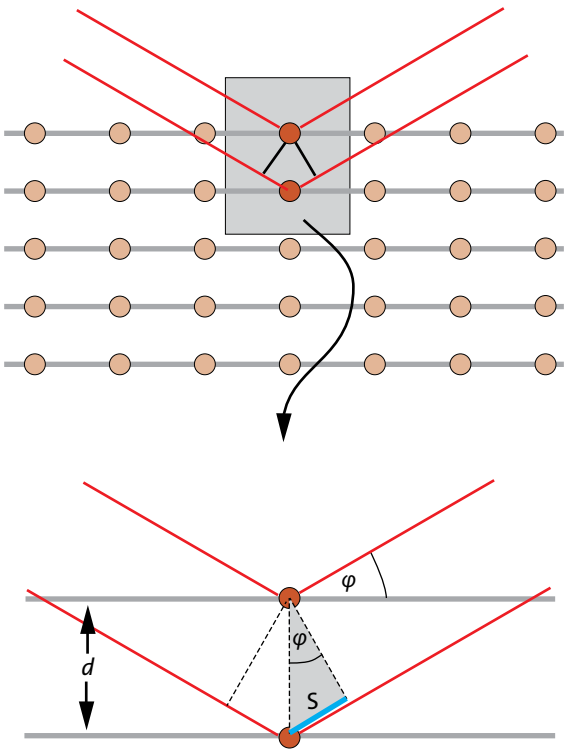

Fig. 5.26 The path of the wave that is reflected on the lower plane is 2*s* longer than the path of the wave that is reflected on the upper plane.

angle of reflection) because the paths of the wave are all equal for this direction.

But behind this first plane of diffraction centers, there is the next plane on which another amplified wave with the same direction as the first one is formed, and on the third plane there will be another one, and so forth. These amplified waves, however, travel paths of different lengths, i.e. they interfere with one another. There will only be amplification with lots of luck: if the paths of the waves that are associated with the different planes differ from each other by exactly λ. And this will generally not be the case. But we can give this process a helping hand: the path difference depends on the angle at which the incident wave falls onto the planes. If we rotate the grating, the waves that are reflected on two neighboring planes will at some point have a path difference of exactly λ, Fig. 5.24. When this occurs, all these waves will amplify and a reflected wave with a high intensity results.

We could have just as well raised the same argument if we had not examined the planes of Fig. 5.24 but any other set of parallel planes on which diffraction centers are located. Fig. 5.25 shows four of many possibilities. For each of these sets of planes we have: if the path difference of waves that are diffracted on neighboring planes is equal to λ, an intensity maximum will emerge. And there will also be a maximum if the path difference is 2λ or 3λ etc.

Hence, we also obtain an interference picture with a three-dimensional grating, but only after rotating the grating around one of its axes. We can calculate the distances of the diffraction planes from the position of the intensity maxima. Now we will see how this can be done.

Planes on which diffraction centers are located are drawn into the grating from Fig. 5.26. We compare the path of the wave that is „reflected“ on the first plane to the path of the wave that is „reflected“on the second plane.

The path difference is $2 \cdot s$, i.e. twice the side s of the triangle highlighted in gray. The relationship between s and the distance d between the planes and the angle

of incidence of the wave can be derived from the drawing. We obtain

$$\sin \varphi = \frac{s}{d}$$

or

$$s = d \cdot \sin \varphi.$$

We obtain amplification when the path difference is

$$2s = k \cdot \lambda,$$

with $k = 1, 2, 3, \ldots$

Amplification in the case of a three-dimensional grating:
$2d \cdot \sin \varphi = k \cdot \lambda$ with $k = 1, 2, 3, \ldots$

If the wavelength is known, we can calculate the distance d of the planes based on an observed angle. To different sets of planes correspond different maxima.

Exercises

1. We let a plane wave, which is composed of sine waves with many different wavelengths from a large λ range, fall onto a three-dimensional grating. (The grating is not rotated.) What can be observed?
2. A plane sine wave falls onto a grating like that from Fig. 5.24. The grating is rotated slowly. Most of the waves moves straightly through the grating. But for certain orientations of the grating, outcoming waves of other directions are formed. Which information about the grating can be obtained from the wave with the slightest deviation from the straightforward direction?

5.7 Diffraction of X-rays in crystals

To obtain interference patterns through diffraction of sine waves, we need gratings whose distance between the slits has similar dimensions as the wavelength.

A diffraction grating for visible light therefore has a slit distance of approximately 0.5 to 1 µm. The wavelength of X-rays is approximately 1000 times shorter. To create an interference pattern with X-rays, a grating with distances between the slits of around 1 nm is needed. Now, we do not even have to deal with the production of such gratings because they are abundant in nature. The atoms of most solid substances are arranged in a regular three-dimensional „crystal lattice". Most rocks and minerals as well as almost all metals belong to these *crystalline* substances.

In some solid substances, the atoms are arranged irregularly. Such substances are referred to as *amorphous*. The amorphous substances include most organic substances in our environment – natural ones such as wood or artificial ones such as organic plastics –, but also glass (in contrast to what you might have expected). In the following, we will examine only crystalline substances, i.e. those with a regular, periodic arrangement of the atoms.

A beam of X-rays with a single wavelength is sent onto such a crystal. The radiation is diffracted on each of the atoms. As we have seen, however, the diffracted radiation interferes in such a way, that the partial waves extinguish mutually so that the X-ray beam simply traverses the crystal in a straight direction. However, if the crystal is rotated, the amplification requirement will be fulfilled for specific directions and a part of the radiation will be diffracted in a well-defined direction.

Fig. 5.27 shows how the examination method works in principle. A thin X-ray beam with a single wavelength, i.e. an X-ray sine wave, comes from the left. It falls onto the crystal to be examined. The crystal is rotated slowly around an axis. Around the crystal there is a photographic film.

During rotation of the crystal, the amplification requirement is fulfilled every now and then so that beams in a variety of directions arise and disappear again. Each of these beams creates a small spot on the film. Hence, we obtain an image with many of such

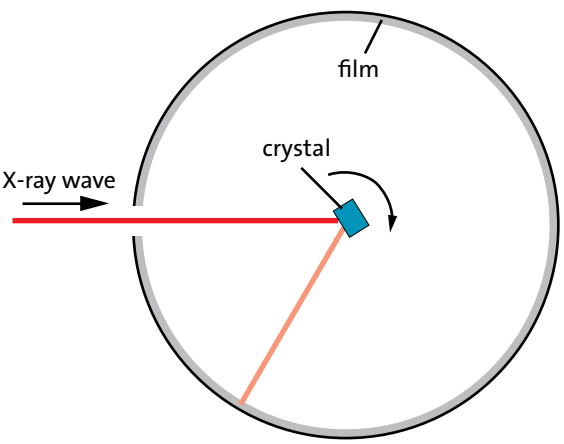

Fig. 5.27 Recording of an X-ray diffraction picture. The crystal to be examined is rotated. The interference maxima are recorded with the cylindrical film.

spots in the end. The structure of the crystal lattice can be calculated from the position and intensity of these spots. The diffraction picture from Fig. 5.28 was created with a gypsum crystal ($CaSO_4 \cdot 2\,H_2O$).

It will become clear how interesting this method is if we consider that there is a very large number of different crystal structures. The interference picture contains this structure in an encoded form. From such pictures one can derive the coordinates of each atom of a complicated molecular crystal.

In fact, this *X-ray structure analysis* provides even more far-reaching information about the structure of matter. While explaining the method, we had supposed that the atoms were punctiform objects on which the sine wave is diffracted.

However, the X-ray wave is not diffracted on the atomic nuclei but on the *electronium*, i.e. the „substance" that is located between the nuclei and whose elementary portions are the electrons. The X-ray waves are diffracted strongest where the electronium density is highest. A detailed analysis of the interference picture does not only allow us to determine the position of the atomic nuclei, but also the density distribution of the electronium. Fig. 5.29 shows such a picture for diamond, i.e. crystalline carbon. At the dark spots (high electronium density), the illustration plane intersects with covalent bonds that have to be imagined as perpendicular to the drawing plane.

Now we can also understand that it is interesting in some cases to use waves of a different nature than X-rays. Depending on the way in which the waves are diffracted in the material to be examined, we obtain a different picture and therefore other information.

For example an *electron beam* is sometimes used instead of an X-ray wave. Electron beams can be created in a way that they form a pure sine wave. Also with such an electron sine wave, important data can be found about the structure of a material. As very thin electron beams can be produced, very small areas of homogeneous material can be examined separately by means of electrons.

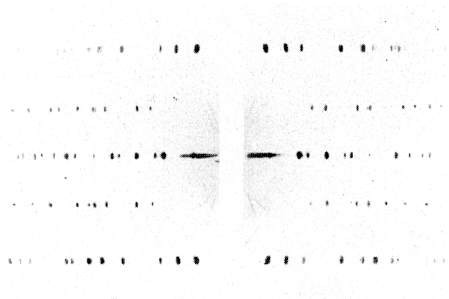

Fig. 5.28 X-ray diffraction picture of a gypsum crystal

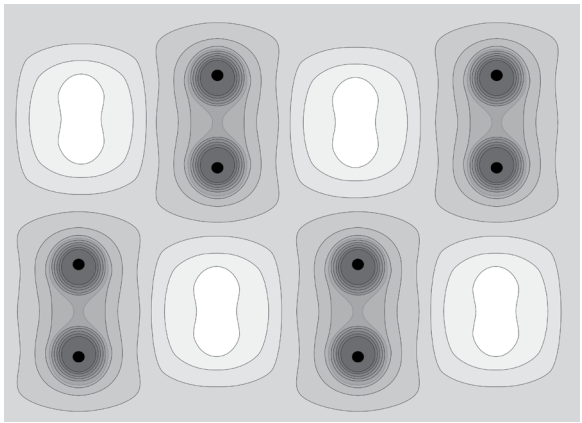

Fig. 5.29 Electronium density of diamond (crystalline carbon) in a plane that does not pass through the atomic nuclei. The density is high (dark) where the plane intersects with the covalent bonds between two adjacent atoms.

Another variant is interference with neutrons. Neutron beams can form a sine wave as well. Neutrons are not electrically charged. Therefore, they are almost not diffracted by the electronium but mostly by the atomic nuclei. One thus can obtain information about the position of the nuclei and about their thermal movement. As neutrons are magnetic, a neutron interference picture also contains information about the distribution of the magnetism in the examined crystals.

6 DATA TRANSFER AND STORAGE

6.1 The amount of data

The German telephone operator (Telekom) establishes telephone connections, a cable network operator provides us with television programs, a provider gives us access to the Internet. All these companies make money by transporting and storing data for other people. They do not care about the nature of the news and information, texts, images or music they transmit for us. From essential news or trivial chatting via telephone to soap operas or reports about a famine via television channels up to a train information or a horoscope from the Internet – there is only one important aspect for the companies: the amount of data that is transferred. The expenditure they make for us and the corresponding fees they charge us at the end of the month depend on this amount. Hence, the *amount of data* is the relevant factor.

The symbol of the amount of data is H, the measurement unit is the bit. The eightfold amount of a bit, i.e. the byte – abbreviated B – is often used. Hence:

1 B = 8 bits.

As very large amounts of data are handled frequently, both the bit as well as the byte are used with the well-known multiplying factors kilo, mega, giga, tera, etc.

How can we determine the amount of data? How much is 1 bit? We could already look at a formula that is used to calculate the amount of data. However, we cannot really tell from this formula why it supplies the amount of data. It is therefore better to consider carefully what we want. Then, it will be easy for us to find the formula ourselves.

In the following, we will examine processes in which data are transmitted or transferred: from a place that we call *data source* to another one, i.e. the *data receiver*. Data transfer is only possible when an agreement has been made between the data source and the data receiver in advance about the signals and signs to be used, i.e. a type of language or alphabet. We call these signs „character set“ and denominate its number with z.

We start with the simplest situation that we can think of: it has been agreed between the data source and the data receiver that only two signs will be used; hence, we have $z = 2$. We can also say that a *binary code* is being used (*Binarius* [Latin] means „consisting of two parts“.)

The nature of the signs is not relevant for our considerations. We could use:

- the spoken words „yes“ and„no“,
- a red and a green light signal,
- a positive and a negative electric potential in a wire,
- we turn our thumb upwards or downwards.

A message that is transmitted by means of such a signal has the amount of data 1 bit. We therefore have a definition of the measurement unit 1 bit, and consequently of the amount of data:

> When the number of signs is $z = 2$, 1 bit is transmitted with one sign.

You bet with a classmate X that you will score a grade A in the next physics test. You get the test back and you wish to tell X, who is sitting on the other end of the classroom, immediately whether you achieved your goal. You transmit the information with your thumb. The transmitted amount of data is 1 bit.

But what is the amount of data for the case that the number of signs is greater than two?

Let's assume we have $z = 4$ different signs. One possibility could be that you wish to tell X the grade (A, B, C or D) you have obtained. (Grade E was excluded from the start.) First we replace the four lettes A, B, C and D by the numbers 1, 2, 3 and 4.

Again, not the realization is relevant for the calculation of the amount of data. You have agreed that you

would hold up one of 4 color pencils. The following *code* was agreed upon:

Grade	*color*
1	red
2	green
3	yellow
4	blue

How many bits are transmitted with the color signal here? We can determine the amount of data by using our first rule. Instead of the four-color code, we transmit the same data with a binary code, i.e. with a thumbs-up and thumbs-down sign. How does that work? You make two successive data transmissions. With the first, you communicate whether the grade is an even or an odd number, and with the second, whether it is the higher or the lower grade value.

Signal 1

Grade	*Thumb*
1 or 3	up
2 or 4	down

Signal 2

Grade	*Thumb*
1 or 2	up
3 or 4	down

The following table shows the two thumb orientations for the four different grades:

Grade	*Thumb*
1	up-up
2	down-up
3	up-down
4	down-down

Let's assume you scored a grade B, or 2 as a number. In this case, you first turn your thumb downwards and subsequently upwards. As the data transmission was now done with a binary code, we know the amount of data: you transmitted twice one bit, i.e. 2 bits in total. Fig. 6.1 shows the *decision tree* for our example. Prior to the transmission of the first signal, all four grades are possible for the data receiver X. With every binary signal, this number is halved, i.e. reduced to two, by the first signal and to one by the second signal. If you had used the four-sign code, you would also have transmitted 2 bits, but by means of a single signal:

If you had used the four-sign code, you would also have transmitted 2 bits, but by means of a single signal:

> When the number of signs is $z = 4$, 2 bits are transmitted with one sign.

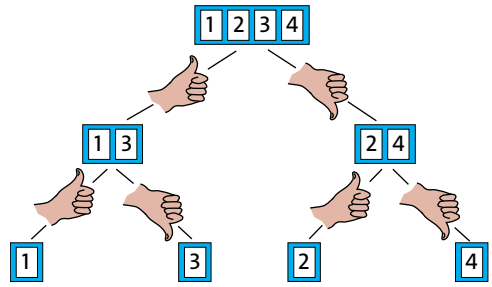

Fig. 6.1 Decision tree for two binary decisions

Hence:

2 different signs: 1 bit/sign
4 different signs: 2 bit/sign

What will be next?

If we wish to transmit a message that consists of a choice from among 8 different possibilities, 3 binary decisions, i.e. decisions between two possibilities, will be required. Three binary signs have to be transmitted. Fig. 6.2 shows the decision tree.

If the message is transmitted with a code that uses 8 signs, we will only need to transmit a single one of such signs. Hence:

> When the number of signs is z = 8, 3 bits are transmitted with one sign.

It goes on accordingly: for z = 16, 4 bits are transmitted with one sign; for z = 32, 5 bits are transmitted, etc., see Table 6.1.

Now we have reached a point where we can get a clear idea of the amount of data. Put yourself in the

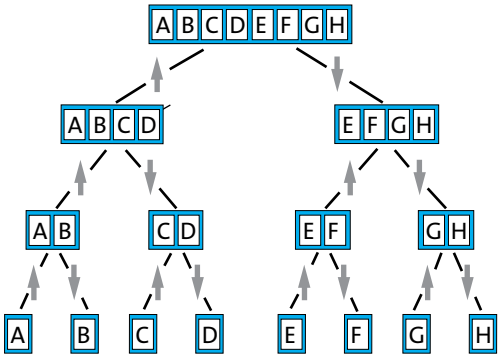

Fig. 6.2 Decision tree for three binary decisions. By means of three binary decisions, one of the 8 signs A, B, C, D, E, F, G or H is selected.

position of the data receiver. What would be your chance to predict a message that has not yet been received?

If it is a 1-bit message, i.e. for example the information whether team A or team B won a soccer game, you will predict the result correctly with a probability of 0.5 or of 50%. (We assume that the teams play equally well.) If the winner shall be determined among 4 teams in a tournament, you will receive 2 bits with the announcement of the winner. Your chance of predicting the winning team was 0.25 or 25%.

Someone thinks of a number from 1 to 64. How likely will you make the right guess if you can name a number? Your uncertainty is much higher now. The probability to make the right guess at the first go is 1/64. If you are subsequently told the correct number, you will receive the amount of data $H = 6$ bit (see Table 6.1).

z	H in bit
2	1
4	2
8	3
16	4
32	5
64	6
128	7
256	8
512	9
1024	10
2048	11
4096	12
8192	13

Table 6.1 Amount of data H for various numbers of signs z

> The higher the amount of data of a message, the more uncertainty is eliminated when the message is received.

By means of Table 6.1, we can determine the amount of data H for a sign out of the number of signs z. You see that the table can be replaced by a simple formula:

$$H = \log_2 z \text{ bit}.$$

In words: the number of bits turns out to be the binary logarithm of the number of signs. As we will need the binary logarithm frequently in the following, we use a special symbol:

$$\log_2 = \text{ld},$$

(„Logarithmus dualis“). We therefore have

$$H = \text{ld}\, z \text{ bit}. \qquad (6.1)$$

Although the formula is very simple, there is a snag: it only applies under specific conditions, and we will only learn about these conditions later. Therefore, be a little skeptical about everything we calculate with this formula now.

There are only powers of two in the left column of Table 6.1. With equation (6.1) we can even calculate the amount of data if z is not a power of two.

Let's assume a test grade should be transmitted once again and that the whole grading scale is possible this time, i.e. one of the grades from A to E. Hence, $z = 5$ applies now. With equation (6.1) we obtain:

$$H = \text{ld } 5 \text{ bit} \approx 2.3 \text{ bit}.$$

We look once again at Table 6.1. It starts with the value $z = 2$ for the number of signs. But what will be the amount of data if only one sign is available for the transmission? We apply equation (6.1):

$$H = \text{ld } 1 \text{ bit} = 0 \text{ bit},$$

because the binary logarithm of 1 is 0. Thus, the amount of data is 0 bits. Can we understand this? Yes, we can. What is the degree of uncertainty that is eliminated by the transmission of the sign? As only one sign is available, the data receiver knows with a certainty of 100 % which sign will come, i.e. the only sign that exists. No uncertainty is eliminated at all. Hence, the formula supplies the expected result also in this case.

But what is the situation in the following cases: the telephone rings, the bell sound at school can be heard or a car honks. It seems that there is only a single sign in each case, but still, data is transmitted without any doubt. But there are actually always two signs: in case of the telephone, the options are „it rings“ and „it does not ring“, in case of the horn it „honks“ or „does not honk“, etc..

When we transmit data from a source A to a receiver B over a longer time, a *data current* flows from A to B. The data current strength (or in short: the data current) is the quotient of the transmitted amount of data ΔH and the time required for the transmission Δt :

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

6.2 Examples for amounts of data and data currents

Writing

Writing is one of the most important methods of storing and transporting data. How many bits are carried by one character? At first, we have to determine the number of existing characters: upper- and lower-case letters, numerals, punctuation marks, math signs and other special characters. Also the space between two words is a character. We assume that only the characters that can be created with a normal keyboard may be used. The keyboard of a typical computer has approximately 50 keys. Each key has a dual function, i.e. depending on whether we press the shift key or not, a different character is written. In case of the letter keys, each one can be used to write either the upper- or the lower-case letter. Hence, a total of around 100 characters can be written with a keyboard. With our provisional formula for the amount of data we obtain

$H \approx 7$ bit.

Images

A computer creates an image on its screen. Which amount of data does the computer send to the screen for this purpose?

In a typical screen, each image point, i.e. each „pixel“ can take on one of 16.7 million different colors and degrees of brightness. Therefore, the amount of data for a pixel becomes:

$H = \mathrm{ld}\,(16\,700\,000)$ bit = 24 bit.

Now we assume the screen to have $4000 \cdot 2500 = 10\,000\,000$ pixels ≈ 10 megapixels. We therefore obtain as an amount of data for the entire picture

$10\,000\,000 \cdot 24$ bit = 240 000 000 bit = 240 Mbit.

This is also the amount of data that is needed for every picture of a digital camera (assuming that the camera takes pictures with 10 megapixels.)

> A typical picture has an amount of data of approximately 200 Mbit.

When an image is saved, it is usually „compressed“. Sometimes, an image that initially contains 200 Mbit can be compressed to 200 kbit, i.e. to a thousandth of the initial amount of data. It will then require less storage space. The fact that nothing of the image is lost in the process appears miraculous at first. You will learn how the trick works in sections 6.4 and 6.5. At the moment, we are only interested in uncompressed data.

„Moving images“ as can be seen on the television screen or at the movies can be obtained by showing or projecting many individual images onto the screen in rapid succession. The movements will appear continuous if more than approximately 20 images are created per second. This resulting data current that has to flow to the projector and to the screen is:

I_H = 240 Mbit per frame · 20 frames per second
= 4800 Mbit/s

> A data current of approximately 5,000 Mbit/s corresponds to moving images.

However, we will see later that a much weaker data current is also sufficient in this case.

Measuring values

When performing a measurement, we receive data about the object on which the measurement is applied.

We examine a meter with an *analog display*. The result is not displayed in numbers, but by means of a pointer on a continuous scale. A kitchen scale, Fig. 6.3, is a typical example.

The maximum value that it can display is 2 kg. The smallest difference that can still be read more or less reliably is 10 g. To the question of how heavy the object is, the scales can therefore give 200 different answers. Thus, the number of signs is 200 and the amount of data that comes with the answer is approximately 8 bit. This value is typical for all meters with an analog scale. Digital meters can have a much higher accuracy. We memorize as an approximate rule:

> An amount of data of approximately 10 bit corresponds to a measuring value.

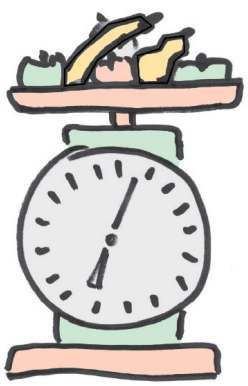

Fig. 6.3 To the question „how heavy are the fruits?“, the scale can give around 200 different answers. The amount of data of an answer is therefore approximately 8 bits.

Music, speech, noise

Data is transmitted as soon as someone speaks or when the radio is on. The transmission occurs by means of sound waves. What is the data current in this case?

We know that the highest perceptible frequency is approximately 20 kHz. To be on the safe side, we calculate with 25 kHz. This corresponds to a period of 0.04 ms (why?). The velocity at which the fastest changes of a sound signal occur is therefore determined. Fig. 6.4 shows the pressure of a sound wave that was recorded with a microphone as a function of time.

To *digitalize* this curve, i.e. if we wish to describe this curve by means of numbers, we have to measure the signal in short intervals. To prevent speech or music from quality losses, the time lags between the individual measurements have to be sufficiently short. To be able to still perceive the fastest occurring changes, two measurements per period of the fastest occurring oscillation are enough, i.e. approximately one value per 0.02 ms or 50,000 per second.

For the transmission to be good, the individual measuring values have to be determined and saved with a high accuracy. We calculate with 10 bit per measurement value. For a stereo transmission, two such measurement values are needed for each instant of measurement.

This leads us to the data current:

$I_H = 2 \cdot 10$ bit per value $\cdot$ 50 000 values per second
= 1 Mbit/s.

> A data current of approximately 1 Mbit/s corresponds to spoken language and music.

We can memorize as a general rule:

> For the transmission of images, the data current is approximately 5,000 times larger than for the transmission of speech and music.

Brain and DNA

Biological evolution has created two „data storages“ that have not yet been outperformed by technical data storage systems in certain respects: the brain and the DNA.

The functioning of human and animal brains has not yet been completely deciphered. We can therefore only make a rough estimate of the data storage capacity. The human brain can store an amount of data of approximately 10^{12} bit.

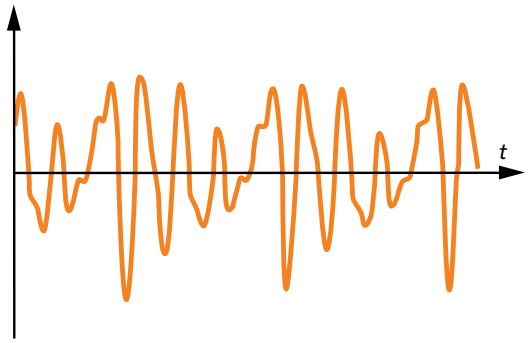

Fig. 6.4 Sound signal recorded with a microphone

A living being comes into existence through the complicated interplay of many chemical reactions. A type of blueprint inherent in every living being ensures the right course of these reactions. This blueprint consists of some very large molecules, i.e. the deoxyribonucleic acid molecules or, in short, DNA molecules. Many of these molecules are contained in every living being, more precisely in the nucleus of every single cell.

The structure of the DNA is easy to describe, Fig. 6.5: between two very long uniform molecule strands, atomic groups that exist as two different types are arranged like rungs of a ladder. Each of these groups can be integrated in the ladder in two ways. Therefore, there are four different „signs“ by means of which the data of the biological blueprint are encoded. As the number of signs is 4, each sign carries 2 bit.

An important goal of biological research consists of deciphering the rules according to which the blueprint of a living being is encoded in the DNA molecules.

Compared to the diameter, the length of such a molecular ladder is incredibly long. While the diameter is approximately 1 nanometer, its lengths is around 1 millimeter, i.e. a million times larger, in a bacterium. The ratio of diameter to length is therefore approximately the same as for a thread with a length of 100 m. In order to fit into the nucleus, the DNA molecule is wound up to a ball.

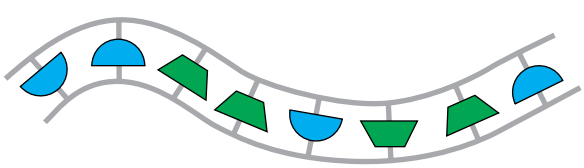

Fig. 6.5 Structure of the DNA, schematic display

The DNA chain is longer the higher the degree of development, i.e. the complexity, of the respective living being. In case of sophisticated organisms, the data is distributed over several DNA clusters, for example over 46 in human beings: This corresponds to a total length of the DNA ladder of 99 cm. The number of rungs in a bacterium is approximately 4 million, and 2.9 billion in a human being. The blueprint of a bacterium therefore has 8 Mbit and that of a human being approximately 6,000 Mbit.

Exercises

1. In the area of Deutsche Post (German mail service), 100,000 different post codes can be used. What is the amount of data that is carried by a post code?
2. The amount of data of a telephone number depends on whether the number is selected from a local network, the national network or the international network. Estimate the amount of data of a telephone number from a local network with 10,000 telephone connections.
3. The Chinese alphabet has many different characters. Approximately 2,000 are sufficient for everyday use. How many bits are carried by one character?
4. A source emits 5 bits with each sign. What is the number of signs of the source?
5. The number of signs of a source is 3. Draw a decision tree for this source. It should comprise three successive decisions. What is the amount of data that a recipient receives with three successive signs from this source?
6. Source A has a number of signs that is a power of two. The number of signs of source B is twice as high. What can be concluded for the amounts of data that both sources emit per sign?
7. A magic trick with cards: We use 16 different cards of any card game. The magician lets a participant draw a card. The participant looks at the card in a way that the magician cannot see it. The card is put back into the card game and the cards are shuffled. Then, the magician reveals the cards one by one by putting them on four different piles: one card on the first, the next on the second pile, one on the third, one on the fourth pile and once again one on the first etc. until all 16 cards are lying on the table. The participant then has to say on which of the four piles his card is lying. Subsequently, the magician makes a new package of the four piles and spreads out the cards in four piles once again and the participant tells him again on which pile his card is lying. The magician now knows the card that the participant has memorized: he merges the four piles again and then flips the cards over one by one until he arrives at the card that the participant has memorized. Which amount of data does the magician has to receive in order to identify one out of 16 cards? How many bits does he get each time the participant names the pile in which the card is located? How does the trick work?

Exercises

8. Estimate how many bits we get if we determine the weight with a balance scale. (Maximum load: 5 kg, smallest weight piece: 1 g.)
9. Look for the amount of data of different files on a computer. What is the type of the files that have a large amount of data? What are the files with a small amount of data?
10. When loading files, the web browser indicates the data current. Take note of some typical values.
11. There is a game in which small colored squared plastic discs are put onto a grating in a way that the grating is covered completely by the small discs. A picture can be composed this way. We assume the grating with a size of 30 cm × 40 cm to have 60 × 80 grid points (i.e. one small color disc has a size of 0.5 cm × 0.5 cm) and there to be discs in 16 different colors. Calculate the amount of data of a picture in two different ways: (a) Calculate the amount of data of one small disc and multiply by the number of the „pixels". (b) Think of each of the different pictures that can be produced as a sign. Calculate how many different pictures there are in total. Calculate the amount of data out of this „number of signs". Compare to the result from part (a).
12. Keys are data carriers. Estimate how many bits are carried by your house key.
13. A music box has 18 tines that can each create one sound. During one rotation of the roller of the music box, one tine can be plucked 20 times as a maximum. How many bits are stored on the roller?
14. Retrieve the information requested in the following from the Internet. (a) Texts have been stored since the invention of a script. Outline the most important development steps of text storage from the beginnings up to today. (b) Outline the most important development steps of the data memory „picture" from the beginnings up to today. (c) Outline the most important development steps of acoustic data stores.
15. Look for the size of the main memory of a computer and check how much of it is utilized. Also try to find out which amount of data fits into the hard drive and how much of it is utilized. Find out the data memory requirement of three different programs.

6.3 Data carriers

We need a *data carrier* to move data from one place to another.

Data carriers for people

Humans receive the largest part of and the most important data through eyes and ears. The data carriers are the light, i.e. electromagnetic waves with wavelengths between 400 and 800 nm, and the sound, i.e. mechanical waves in the air with frequencies between 20 Hz and 20 kHz. (For light, the wavelength can be

measured more easily, and the frequency for sound. This is the reason for the different measurement units.)

Technical data carriers

Electricity

When an electric connection is used, the data carrier is electricity, for example:

- the cable between the amplifier and the speakers
- the cable from the bell push to the bell
- the telephone line
- connections within electronic devices
- the cable from the antenna to the television.

Electromagnetic waves

They are so important that they deserve a special treatment; see sections 6.9 to 6.11. Also the light is an electromagnetic wave. For technical use, it is not relevant whether the waves belong to the visible light or not.

Mechanical data carriers

They have practically disappeared from use although they were very significant in earlier times. For example railway signals were operated remotely by means of a cable-pull. Also doorbells worked by means of bars and ropes.

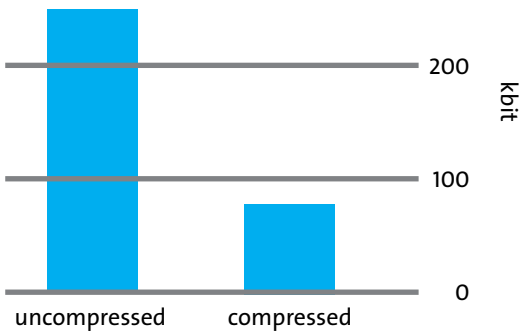

Fig. 6.6 Uncompressed (= redundant) data and compressed data

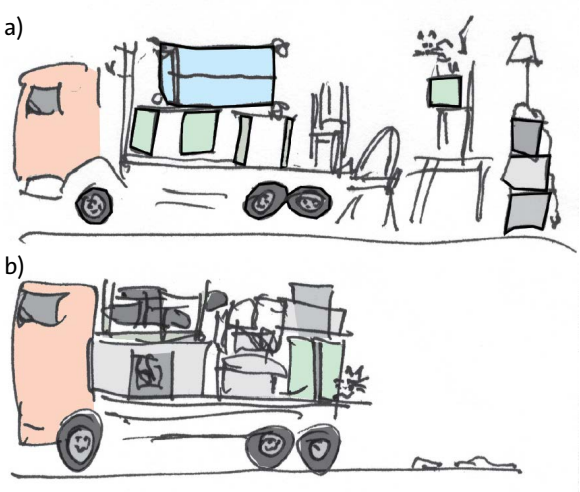

Fig. 6.7 (a) Uncompressed removal material; (b) compressed removal material

6.4 Actual and apparent amount of data

We have learned how to calculate the amount of data. It is equal to the binary logarithm of the number of signs:

$H = \mathrm{ld}\, z \text{ bit}$.

But we have already noticed that the formula is not yet complete. When applying it, something must be taken into account – and this will be explained now.

Indeed, it is possible that a specific amount of data – let's say 80 kbit – is encoded unhandily so that it takes up more storage space, for example 250 kbit. We therefore distinguish between the true amount of data, i.e. 80 kbit, and the apparent amount of data, i.e. 250 kbit, Fig. 6.6.

As we mostly deal with the apparent amount of data, the adjective „apparent" is mostly omitted in the following. Hence, „amount of data" (without any adjective) means apparent amount of data. In our case, we can say:

- actual amount of data: $H_0 = 80$ kbit
- (apparent) amount of data: $H = 250$ kbit

The amount of data can be reduced through re-coding until it is finally equal to the actual amount of data. We say that the data is *compressed*. As long as $H > H_0$ applies, the data is *redundant*. Hence:

> Redundant data can be compressed.

We explain the situation by means of an analogy:

A furniture truck has a loading capacity of 40 m^3. Now, furniture and other removal material is loaded, Fig. 6.7.

Before, we determine the volume of the removal material by calculating length times width times height

for each object: for tables, cupboards, chairs and all the other furniture, for cardboard boxes and any other stuff. We find 110 m^3. The moving truck would consequently have to drive three times to take everything away. But we can immediately see that more things can be stored by skillfully packing things into each other. For example an entire chair can be stored between the legs of a table. And there are many other smaller gaps that can be filled with cardboard boxes and other small objects. Finally, a lot can be stored in drawers of closets, cupboards and desks. In the end, everything fits into the storage space of a single truck, and there is still a lot of space left. We have „compressed" the removal material from 110 m^3 to 36 m^3. A very similar procedure can also be used for the data. Poorly packed, they need 250 kbit; compressed only 80 kbit.

We had formulated the rule:

„*The larger the amount of data of a message, the more uncertainty is eliminated upon arrival of the message.*"

Here, we mean the actual amount of data because only the portion H_0 of the amount of data eliminates uncertainty. What goes beyond that is responsible for redundancy and will not eliminate any uncertainty at the data receiver. We can therefore reformulate:

> The higher the actual amount of data of a message, the more uncertainty is eliminated during arrival of the message.

Of course, we are very interested in reducing the „storage space" that the data is taking up. We would like to compress the data as far as possible, i.e. until the (apparent) amount of data has become equal to the actual amount of data:

$$H = H_0$$

In case of furniture, we can tell relatively easily whether the pieces can be nested into each other even better. Can we also tell whether data are still redundant, i.e. whether they can be compressed further? We can tell indeed.

> Data are no longer redundant when two preconditions are met:
> - The probabilities of all signs are equal.
> - The probability of a sign is independent of the preceding signs.

Now we can understand why the equation

$$H = \text{ld}\, z \text{ bit}$$

has to be applied with caution. It does not supply the actual but an apparent amount of data. It tells us how much memory space is needed, regardless of whether the data is compressed or not. Only if the preconditions 1 and 2 are met, it will provide the actual amount of data.

6.5 The principle of data compression

Most data we are dealing with are at first available in an uncompressed, i.e. redundant, form. Much more memory space is needed than what corresponds to the actual amount of data. Earlier, we estimated the following values for the amounts of uncompressed data and data currents:

- picture: $H \approx 200$ Mbit
- moving images: $I_H \approx 5000$ Mbit/s
- spoken words and music: $I_H \approx 1$ Mbit/s

It is not unusual for the apparent amount of data to be 1000 times as large as the actual amount of data. Therefore, compression of data is a profitable business.

How do compression methods work. We look at two simple examples that illustrate the principle.

1. Redundancy through conditional probabilities

We want to transmit an image whose pixels are only black or white like in the case of old fax machines, Fig. 6.8.

At first, we save the data in the simplest possible way. Each pixel corresponds to one of two possible signs, i.e. one for black and the other one for white, for example:

- white 0
- black 1

The whole picture will then be encoded by indicating for each line and from the top left to the bottom right – pixel by pixel – whether the corresponding pixel is black or white.

We limit ourselves to the small section that is shown in the enlarged picture at the very bottom. The section contains 40 · 10 pixels = 400 pixels. By means of our code, we obtain a sequence of zeros and ones:

0000000000000000000000000000000000000000
0000000000000000000000000000000000000000
0000000000000001111111110000000001111000000
0000000001111111111111000000001111000000000
0001111111111111111110000001111000000000011

11111 11111111111100000111100000000000111111111111111111111100011110000000000111111111111111111111100111100000000011111111110000000011111111000000000000000111111100000000000000111111111110

We obtain 400 bits as an amount of data. This encoding is redundant because precondition 2 is not met. The probability for the occurrence of each sign depends on the preceding sign. After a zero (for white) a zero will appear with a much higher probability than a one. And after a one, the probability of there being once again a one is much higher than the probability of a zero. The reason: there are large continuous areas in the picture that are completely black and others that are completely white.

Therefore it is not difficult to compress the data. The old binary encoding starts with 101 zeros in a row. Then, there are 9 ones etc..

We can write instead:

101w9b10w4b15w13b8w4b13w17b6w4b12w19b5w4b11w22b3w4b10w24b2w4b9w9b7w9b14w7b12w8b1w

Here, „w“ stands for white and „b“ for black. We have used a total of 81 signs. What is the amount of data now? Our new number of signs is 12 because we use the ten figures 0 to 9 and the two letters b and w. The amount of data per sign therefore results in:

$$H = \mathrm{ld}\ (12)\ \text{bits} \approx 3.6\ \text{bit}.$$

As we have 81 signs, the total amount of data of our image section is

$$H = 81 \cdot 3.6\ \text{bits} \approx 292\ \text{bit},$$

i.e. clearly less than the initial 400 bit.

Of course, the method also works in cases where the pixels take on more than only two colors, i.e. in case of colored images.

But the compression possibilities are not yet exhausted at this point. We have for example not yet taken advantage of the fact that the black and white pixels also form continuous areas in the vertical direction as well. Let's assume we have just arrived at pixel P from Fig. 6.9 with our transmission. This pixel is black with a high probability not only because the previously transmitted pixel Q is black, but also because pixel R, which is located directly above, has already been black.

If we also consider the dependencies in the vertical direction for encoding, redundancy will be decreased even further.

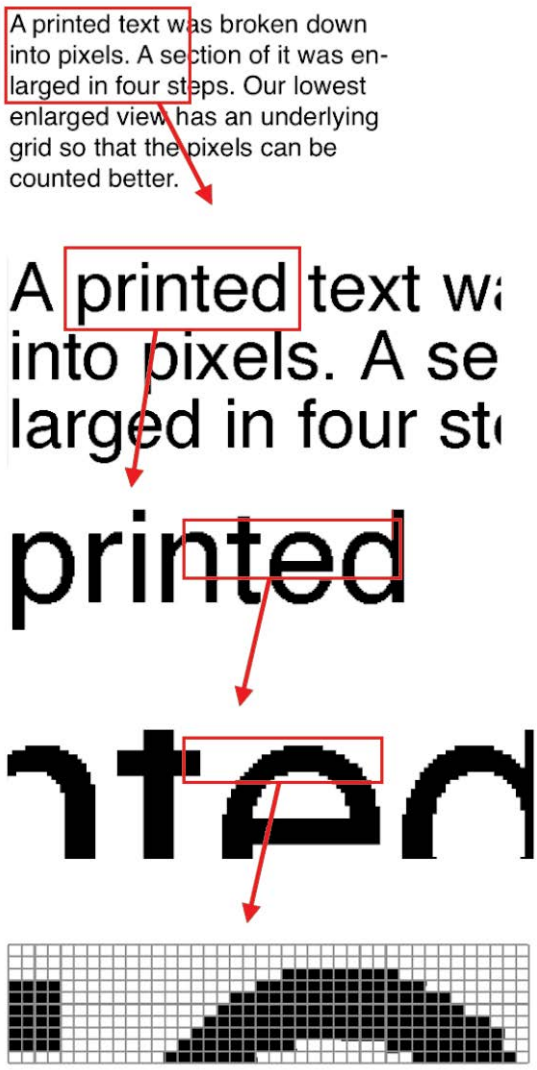

Fig. 6.8 (a) A printed text was broken down into pixels. A section of it was enlarged in four steps. Our lowest enlarged view has an underlying grid so that the pixels can be counted better.

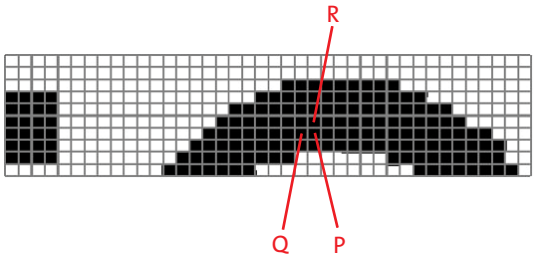

Fig. 6.9 Prior to sending pixel P, we know that its probability of being black is high: 1. because pixel Q is black and 2. because pixel R is black.

		binary code	
sign	probability	without compression	with compression
a	0.6	00	1
b	0.2	01	01
c	0.1	10	001
d	0.1	11	000

Table 6.2

2. Redundancy due to unequal probabilities

A text should be transmitted by means of a letter code. We calculated earlier that approximately 7 bits per letter are needed. But this is an apparent amount of data. Also these data are actually redundant. Why? The probabilities of the different letters and special characters are different. Thus, our first condition (section 6.4) is not fulfilled. The "e" is almost five times as probable as the "u" and around thirteen times as probable as the "v".

We examine with a somehow unrealistic but illustrative example how such signs with different probabilities can be compressed. We assume our alphabet to consist only of the four letters "a", "b", "c" and "d". We suppose the probabilities of the four signs to be the values of the second column of Table 6.2.

The third column of the table contains a normal binary encoding. As there are 4 letters, we have z = 4, and the result is 2 bit per sign for the amount of data. This is also shown by the fact that we need two binary signs for each letter in the binary code. Now we can already see how the data can be compressed. To transmit the frequently occurring „a“, we have used as many binary signs as for the rare „c“. If the encoding is changed in a way as to transmit the frequent letters with fewer binary signs and the rare ones with more binary signs, we can win. A little problem is the condition that we will not know at first where the binary sign sequence, which corresponds to a letter, ends. But due to the way in which the encoding of the fourth column of the table is chosen, the problem does not occur. Any sequence of zeros and ones corresponds to exactly one sequence of letters of our original signs. Hence, we have for example:

0100100000011101010011000 1…
→ bcddaaabbcada…

Now, we need a different amount of bits for the transmission of different signs. We can see whether our encoding has been effective if we calculate the average number of bits: the number of bits for every single letter weighed with its probability.

$$H = \underbrace{0.6 \cdot 1\,\text{bit}}_{\text{a}} + \underbrace{0.2 \cdot 2\,\text{bit}}_{\text{b}} + \underbrace{0.1 \cdot 3\,\text{bit}}_{\text{c}} + \underbrace{0.1 \cdot 3\,\text{bit}}_{\text{d}}$$
$$= 1.6\,\text{bit}$$

The amount of data has decreased as we compressed the data.

6.6 A few frequently used encodings

Images

The file of a normal, uncompressed image has an approximate amount of data of 200 Mbit.

There are numerous compression methods for such files. If you save a picture from a drawing program, the computer will ask you about the „format“ in which you want to save it. Depending on the image type and the purpose of use, one or the other encoding is suited better. The amount of data is different in each of these formats. Here are two examples from the list:

- JPEG: strong compression. Is also used in the digital camera.
- BMP: no compression.

Music, spoken words

A data current of approximately 1 Mbit/s corresponds to uncompressed acoustic data.

On a CD, data are saved as they occur, i.e. in a completely uncompressed way, just like on an old music cassette or a vinyl record.

The data will need much less memory space if they are encoded in the MP3 format. However, the MP3 method has a particularity: not only the apparent, but also the true amount of data will be reduced. Some details of the original sound file are simply omitted. This can be done because humans do not perceive the difference. But it also means that the original sound file cannot be recovered from the MP3-file.

The data current is reduced to approximately 1/10 through MP3 encoding:

Acoustic data
uncompressed: I_H = 1000 kbit/s
MP3-encoded: I_H = 100 kbit/s

Moving images

In an uncompressed form, the data current is approximately 5000 Mbit/s. The respective amount of data was saved on a video tape by means of an old video camera.

But the successive images of a movie sequence are very similar to each other, Fig. 6.10. After transmitting or saving one image, there will be little new information with the next image. Only little uncertainty is eliminated. The data are consequently very redundant.

The MPEG encoding eliminates this redundancy to a large extent. Therefore, the data current is reduced to 1/10. It is consequently not surprising that a 2-hour movie on a DVD only requires 70 Gbits of memory space.

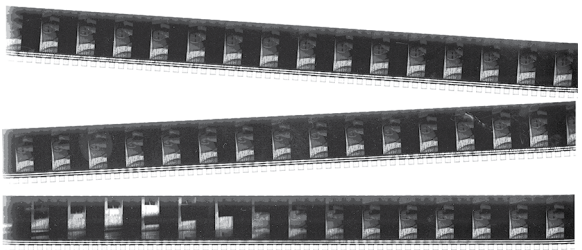

Fig. 6.10 Sequence of an old cinema movie. An image does not contain much new information in relation to the preceding images. The data are very redundant.

Texts, technical drawings and instrumental music

Images and music often consist of elements that are frequently recurring and that the data receiver already knows:

1. A text consists of letters. When it is transmitted by fax, we explain time and again to the data receiver how an „a", a „b", etc. looks like. But the data receiver already knows this. No uncertainty is therefore eliminated in this respect. All the receiver does not yet know is the order in which the letters are arrive. The fax is strongly redundant for this reason.

To eliminate this redundancy, the text is not encoded as an image but as a sequence of letters. We could see earlier that 7 bit per letter are required. The code that has been agreed upon is called ASCII code. In fact, not 7 but 8 or more bits are used per sign. This way, many special characters can be encoded besides the normal alphabetic signs. The ASCII code is used on many occasions, for example:

- to transmit the data from the keyboard to the computer;
- to save the files created with a word-processing program;
- to transmit a text message or e-mail.

2. A technical drawing consists of line segments, rectangles, ellipses, circular arcs, dashed lines, etc.. When the drawing is transmitted as a BMP file or also as a JPEG file, the data receiver is told time and again how an ellipse, a straight line or a rectangle looks like – something that he already knows, though. What he does not know is the length of the line segmens, the size of the ellipses or rectangles, etc., how they are arranged, the thickness of the lines… BMP and JPEG files are therefore redundant.

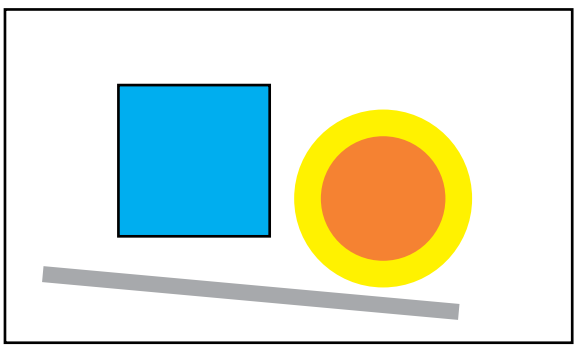

Fig. 6.11 The coordinate zero point for the data of the vector graphic is the top left corner of the figure.

To eliminate this redundancy, one does not encode the drawing as a pixel image but as a *vector graphic*. The „technical drawing" from Fig. 6.11 could for example be described by the following text:

Square
Edge length: 2 cm
Position of top left corner: (1.5 cm; 1.0 cm)
Orientation: 0°
Line thickness: 1 pt
Line color: 0, 0, 0, 100
Filling color: 100, 0, 0, 0

Circle
Radius: 1.0 cm
Position of center: (5.0 cm; 2.5 cm)
Line thickness: 10 pt
Line color: 0, 0, 100,0
Filling color: 0, 60, 90, 0

Straight line
Coordinates of start: (0.5; 3.5)
Coordinates of end: (6.0; 4.0)
Line thickness: 6 pt
Line color: (0, 0, 0, 40)

The amount of data of this text is much lower than that of the pixel file. There are many different vector graphic formats because each drawing program has its own format.

3. A musical piece consists of sounds of known instruments such as clarinet, violin or piano. If the piece is uncompressed or transmitted as an MP3 file, the data receiver will be told what a clarinet, a violin or a piano sounds like although the data receiver has already had this information. No uncertainty is eliminated in this respect. What he does not know is the temporal order of the sounds and which instruments creates a sound at what time, for how long and at which sound volume. The file is redundant for this reason.

To eliminate this redundancy, we can create a *MIDI* file. In this process, the properties of every single sound are encoded: pitch, start and end time of the sound, instrument, sound volume etc. A MIDI file can be created by means of the electric piano and played with a synthesizer.

We have addressed three encoding methods that have something in common. One advantage is obvious: the amount of data is low, i.e. we need little memory space. The methods also come with another benefit that we will explain by means of the example of text files. If we transmit a text with a fax machine and subsequently enlarge it, we can see imperfections. We see the individual pixels and the letters are blurred. When the text is transmitted as a text file, the quality of the letters is only limited by the printer, and this is also the case when they are enlarged.

What are the disadvantages of these storage formats?

Exercises

1. Create a sequence of zeros and ones with a length of approximately 40 signs by tossing a coin. What is the corresponding sequence of the letters a, b, c and d when you interpret it as a message that has been encoded according to Table 6.2 column 4?
2. Messages that are at first available in a code with 8 different signs should be transmitted. Hence, we have: z = 8. We denominate the signs with the first 8 letters of the alphabet. The probabilities at which the signs occur are in brackets: a (0.6); b (0.2); c (0.1); d (0.06); e (0.02); f (0.01); g (0.005); h (0.005).
   (a) Indicate a binary code that does not compress the data. What is the amount of data for each of the 8 signs?
   (b) Look for a binary encoding method through which the data are compressed. What is the mean amount of data for each of the 8 initial signs? Make sure that the sequence of binary signs allows for unambiguous decoding.
3. When a text in a certain language – for example English – is transmitted, also the compressed letter code is still redundant. Which encoding method could be used to reduce redundancy even further?
4. Estimate the amount of data that can be saved on a CD.
5. Create a simple image with a drawing program. Save it in different formats, in particular as a BMP and as a JPEG file. Compare the sizes of the files.
6. Willy and Lilly are sitting with their backs towards other. Lilly has a picture in front of her and describes it with words. Willy draws a new picture according to this description. Try this image transmission method with a friend. Compare the encoding with the methods discussed in the text. Which method is most similar to it? How can it be improved?
7. A musical piece is available in two formats: in MIDI format and in MP3 format. What are the disadvantages of the MIDI encoding?
8. A CD with an amount of data of 480 MB is copied. Will the amount of data double in this process? Explain.

6.7 Games

In some games it is important to ask a question in such a way as to obtain a maximum information with the answer. Hence, the answer should have as little redundancy as possible. We start with a very simple version of such game.

Quizzes

Lilly thinks of one of the integers from 1 to 64. Willy should find out the number by asking Lilly as few as possible yes/no questions.

A „yes/no question“ is a question that can only be answered with either „yes“ or „no“. Therefore the amount of data that is transmitted with the answer to a yes/no question is 1 bit. Whether this amount is the actual or only the apparent amount of data depends on the probabilities of the two answers. If „yes“ and „no“ are equally likely, it will be the actual amount of data. If the two answers have different probabilities, 1 bit is only the apparent amount of data. The actual amount of data will then be lower; the answer is redundant.

We assume Lilly thinks of the number 28. Willy can apply different guessing strategies. We compare two such strategies.

Strategy 1

B: Is it the 1?
A: No.

B: Is it the 2?
A: No.

B: Is it the 3?
A: No.
.....

B: Is it the 28?
A: Yes.

Willy needed 28 questions to find out the number.

Strategy 2

B: Is the number larger than 32?
A: No.

B: Is the number larger than 16?
A: Yes.

B: Is the number larger than 24?
A: Yes .

B: Is the number larger than 28?
A: No.

B: Is the number larger than 26?
A: Yes.

B: Is the number larger than 27?
A: Yes.

Willy knows the number after having asked 6 questions.

The second strategy is obviously better than the first one. There, the questions were formulated in a way that the two answers „yes" and „no" were always equally likely. With each answer the actual amount of data, i.e. 1 bit, was supplied.

For almost all questions of the first strategy, the answer „no" was more likely than the answer „yes". Therefore, the actual amount of data of an answer was smaller than 1 bit.

The fact that Willy receives fewer bits when using the bad strategy than when using the good one is also in agreement with our rule:

„The larger the actual amount of data of a message, the more uncertainty is eliminated upon arrival of the message."

In fact, Willy has a good chance of predicting the answer to each of the many questions he asks when using the bad strategy. He knows that to the question „is it the number 1?", the answer is very likely to be „no". The probability of him being wrong is only 1 : 64, the probability of him being right is 63 : 64. That is why he has to ask a very large number of questions and he only receives a very small amount of data with each answer. In case of the good strategy, there is maximum uncertainty about the next question and the amount of data he gets with each answer is large, i.e. 1 bit.

A slightly more interesting variant of this game is the subject of exercise 2.

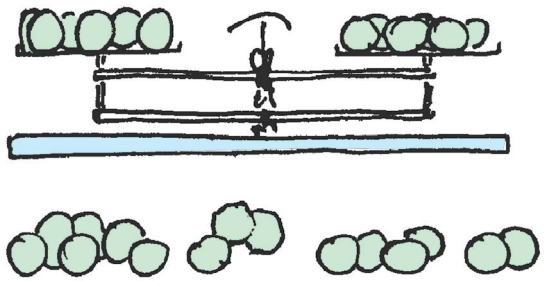

Abb. 6.12 The 27 balls look the same, but one of them is heavier than the remaining 26. How many weighings are required to find the „odd" ball?

The best weighing strategy

One of 27 identical balls is heavier than the remaining 26 ones with an equal weight. We would like to find the heavier ball by means of a balance scale and with as few weighings as possible. Thereby, only balls and not weight pieces or any other bodies may be placed on the weighing pans, Fig. 6.12.

With each weighing step, the balance answers one question it is asked. The balance scale can give three different answers: 1. The right weighing pan goes down, 2. the left weighing pan goes down and 3. equilibrium.

If we seek to perform as few weighings as possible, we will have to ask the questions in a way as to obtain a maximum number of bits per weighing. This means: the probabilities of the three answers have to be as similar as possible for each weighing step. It is certainly awkward to start with placing one ball on each weighing pan. The probability of the scale to remain in equilibrium will then be much higher than the probability of the scale being inclined right- or leftwards.

How many weighing steps are necessary? What is the best strategy?

The following version of the game is substantially more difficult: one of 12 balls has a deviating weight but we do not know whether the „odd" ball is lighter or heavier than the others.

Exercises

1. Willy throws a common dice (i.e. with numbers from 1 to 6). Lilly should find out the number with as few as possible „yes/no questions“. How can Lilly ask the first question in order to obtain 1 bit with the answer? Name two possibilities. Explain why Lilly gets less than 1 bit with the answer to the question „is it the number 6?“.
2. Lilly thinks of a random word. Willy has to find out the word by asking Lilly as few as possible yes/no questions. What is the strategy that Willy has to apply? What is the approximate number of questions that have to be asked in case of this strategy?
3. A mole A presumes that the Sun rises in the west on some days. He hires mole B to make the observation. B goes out every day, looks where the Sun rises – in the east or in the west – and tells A about the result of the observation via a data line. Of course, the data are encoded: one sign for east and another one for west. What is the apparent amount of data? How much uncertainty is eliminated through the transmission? What is the actual amount of data? Explain.
4. You probably know someone who gets on your nerves because he/ she always tells you the same stories. Explain your annoyance by means of a statement about the amount of data.

Fig. 6.13 (a) Unreduced removal material (b) Reduced removal material

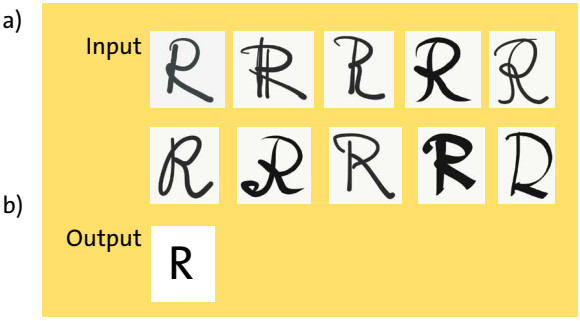

Fig. 6.14 The computer transforms each entered R (a) into the same R (b).

6.8 Data reduction

To save memory space, we are eager to compress data. Compression means: reducing the apparent amount of data while keeping the actual amount of data as it is.

But also the actual amount of data is often reduced. Why? Simply because a part of the data is not interesting. In that case, we talk about *data reduction*.

Explaining it once again with the example of the furniture truck (see section 6.4): out of the 36 m^3 of removal material, 12 m^3 are not even loaded onto the furniture truck but put directly to the bulky waste because the things are no longer useful, Fig. 6.13. We can therefore also say that data reduction consists of throwing away „data waste“.

Data compression:
- apparent amount of data is reduced
- actual amount of data remains constant

Data reduction:
- superfluous information is discarded
- actual amount of data is reduced

Data reduction in case of pattern recognition

Letters can be entered handwritten into the tablet computer. They are at first available for the computer as a pixel file with a large amount of data. A „character recognition“ program converts this pixel file into a text file. The true amount of data decreases very strongly in this process. Of course, much information is lost, too. A letter can be written in countless ways, Fig. 6.14. We can no longer tell from the R, which is created by the computer on the display field, how the handwritten R looked like.

Let's assume that the pixel file on a small screen has an amount of data of 8 kbit. A letter saved in the ASCII format requires 8 bits. Hence, the computer has reduced the amount of data to a thousandth.

A computer can also be programmed in a way that it detects other shapes or „patterns“: squares, circles, straight lines, houses, trees, animals, fingerprints, faces and much more. The true amount of data is reduced in any case.

Pattern recognition is based on data reduction.

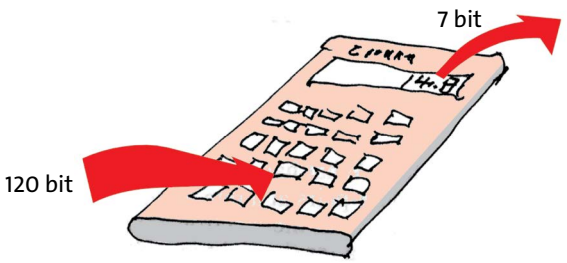

Fig. 6.15 120 bit enter the calculator, and 7 bit come out of it.

Perception and data reduction

What is done by the computer in this case is done by our brain every second. We look out of the window. Immediately after, however, we will no longer remember the complex color pattern that we have perceived with our eyes. We rather know: here is a house, there is a road, there are two people and here is a dog.

The huge data current that enters the eye through the pupil could not be processed further by the brain without reduction.

> Perception is based on data reduction.

Data reduction during calculation

We imagine the calculations to be performed by a computer or a calculator.

A computer receives and releases data. Does this mean that it only re-encodes the data, i.e. that the actual amount of data at the input is the same as at the output?

To see whether this is the case, we look at a simple example. The average of school grades should be calculated. A class with 30 students did a test in which a maximum of 15 points could be achieved. Hence, 30 numbers of which each one is one of the 16 integers from 0 to 15, are typed into the calculator. As $16 = 2^4$, the calculator receives 4 bit with each figure, i.e.

$H_{in} = 30 \cdot 4$ bit = 120 bit in total.

The average value is calculated down to one digit after the decimal point. The result is one of the values

0; 0.1; 0.2; 0.3; ..... 14.7; 14.8; 14.9; 15.

There are 151 possibilities. Therefore the number of signs of the result is $z = 151$. From this the amount of data can be calculated:

$H_{out} = \mathrm{ld}\,(151) \approx 7$ bit.

The calculator has reduced the amount of data from 120 bit to 7 bit, Fig. 6.15.

> Calculators reduce the amount of data.

So does this mean that someone who receives the data of the output knows less than someone who receives the data of the input? Yes, this is exactly the case. Those who only know the average value cannot derive the individual grade points of the students on that basis.

But why is the calculator still being used? It is used precisely because the large amount of data at the input cannot be handled easily. If we wish for example to compare a class as a whole with a parallel class, the amount of data of the individual grades is usually too large. In case of large amounts of data, people easily get confused. Therefore, the calculator is not used because there is too little, but because there is too much data.

Data reduction during recording of acoustic signals

When music or other acoustic data is recorded with a microphone and saved on a CD, there is much data waste among the data. Much of what has been recorded and what is emitted by the speaker cannot be perceived. The microphone absorbs sound waves of all frequencies with roughly the same accuracy. The sensitivity of our auditory sense, however, varies greatly for the different frequency ranges. In addition, the sensitivity of the auditory sense will be reduced very strongly for a wave of a specific frequency if waves of other frequencies are added. These effects can be taken advantage of in order to eliminate useless data. Hence, we can save only what can actually be heard. The method requires a high calculation workload but it is worth the effort. It is the well-known MP3 method. As an MP3 file, a musical piece only requires approximately one tenth of the memory space that it would require in an unreduced state on a CD – without changing the sound quality.

> During encoding in the MP3 format, the amount of data is not only compressed but also reduced.

Data reduction during recording of optical signals

The situation is similar for optical signals. We look at a pixel of a video camera. The light that hits the pixel has a complicated spectrum. If we were to capture

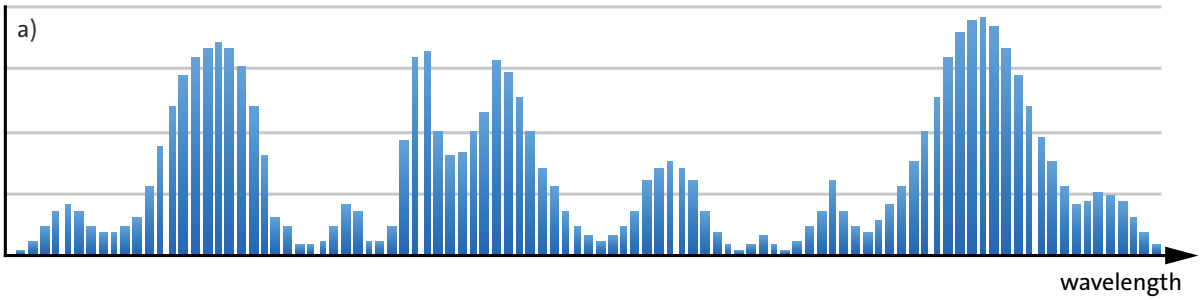

Fig. 6.16 (a) The spectrum is described by 100 values with 8 bit each, i.e. around 800 bit altogether. (b) Only three values with 24 bits in total are required to describe the color impression.

this spectrum with some accuracy, we would have to measure the light intensity for many wavelengths and encode said light intensity as numbers. Assuming that we break down the spectrum into 100 wavelength intervals, we would need around 800 bit for each pixel, Fig. 6.16. Then, the spectrum could be reproduced somehow in the playing device.

In fact, however, a lot of useless data would have been absorbed, transported and reproduced in this case. This is because our eyes cannot distinguish many different spectra from each other. All color impressions that we perceive can be described with only 3 numbers, e.g. a red, a green and a blue value. It is consequently reasonable to reduce the amount of data in a way that the superfluous information does not need to be transmitted. The data reduction is very simple in this case. It is already done by the camera. The camera only records 3 different color signals from the outset.

Video camera and digital camera reduce the spectral signal to three numbers.

Exercises

1. Most computing operations reduce the amount of data; some others leave it unchanged. Name a few examples.
2. Explain why the computer performs a data reduction when it calculates the value of x^2 for an integer x.
3. Will the amount of data be reduced if the value of x^3 is calculated? (x = integer).
4. Willy stayed with his grandparents for one week. Back home, his parents asked him about what the grandparents had told him. Willy's report is rather short. With regard to the grandfather, Willy has compressed the amount of data. And he has reduced grandmother's data.
   What can you say about grandfather's and what about grandmother's narrative style?

6.9 Data transmission with electromagnetic waves – carrier waves

Electromagnetic waves from a very large wavelength range (and/or frequency range) are used: from approximately 500 nm, i.e. from visible light, to 100 km.

We usually aim at making several data transmissions at the same time at some place or in a specific area. You know that it is possible to receive different radio and television channels simultaneously. In addition, wireless phones, smartphones and laptops work at the same place. Radio clocks are controlled and positions are determined by means of GPS (Global Positioning System). Taxis are called over a radio connection and many other services, too. All this occurs by means of electromagnetic waves that run around simultaneously. In every place, waves with diverse directions, amplitudes and wavelengths move through each other. How can we prevent the different systems from disturbing one another?

Everyone who seeks to transmit data is assigned a carrier frequency with a specific, narrow frequency interval on both sides of the carrier frequency. They may only use sine waves from this channel for their data

transmission. The frequency ranges of the different users are very close to each other but none of them disturbs any other one, Fig. 6.17.

To „charge“ data onto the electromagnetic wave, we start from a sine wave with the carrier frequency and modulate this wave. This means that we create small deviations from the sine function and change these deviations in the rhythm of the signal to be transmitted. We will see in the next section how this works in detail.

The stronger the data current to be transmitted, the larger the required frequency interval (in technical terms: the „bandwidth“). Hence, television channels are assigned a frequency range of 7 or 8 MHz, VHF radio channels only 100 kHz.

Of course, the resulting signal is no longer purely sine-shaped. But it can be broken down into sine portions (see chapter 3). Modulation has to ensure that the modulated signal will only contain sine portions from the assigned frequency range. Otherwise, the reception in other channels would be disturbed.

Fig. 6.18 and 6.19 symbolically show the individual components of a data transmission and receiving system: In the transmitting antenna, the signal to be transmitted as well as the sine-shaped carrier signal are fed into the „modulation“ box. The sine function is changed by means of the signal to be transmitted. The respective current flows through the sending antenna and this antenna emits a corresponding electromagnetic wave.

In the receiving antenna, the arriving wave creates an electric current, whose time dependence is the same as that in the transmitting antenna, through electrostatic and electromagnetic induction. But also other waves from other data sources arrive at the receiving antenna. The frequency interval that is associated with the desired data source is filtered out of this muddle („filter“ box). In the „demodulation“ box, the initial signal is distilled back out of the nearly sineshaped current. The desired channel, i.e. the frequency interval associated with the desired data source, can be set on the filter.

This method can be used to filter even a very weak signal out of an extreme chaos of signals with a variety of carrier wavelengths that are sometimes very strong. You can imagine the situation as follows: there is a terrible noise in a room. Hundreds of musical pieces are played at the same time while there are also ugly sounds such as screams, thunder, creaking and squealing noises. Also the lowpitched chirp of a cricket is part of this din. Taking the detour using the carrier frequency, we are able to suppress the overall din in a way that only the chirp will be heard.

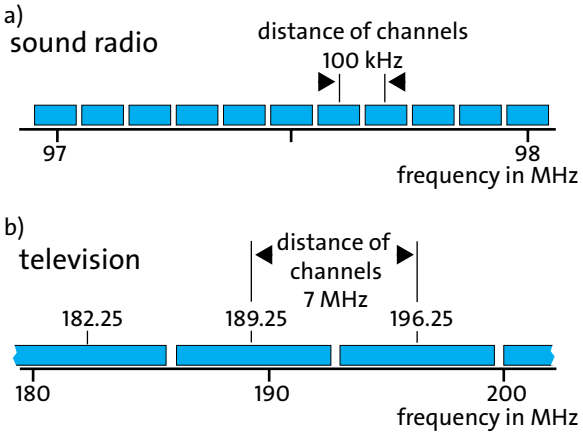

Fig. 6.17 Those who wish to transmit data with electromagnetic waves will be assigned a specific frequency range. The data source may only create frequencies from this range. The channels are illustrated by the gray stripes. Sections from the range of the radio channels (a) and the television channels (b). Mind the different

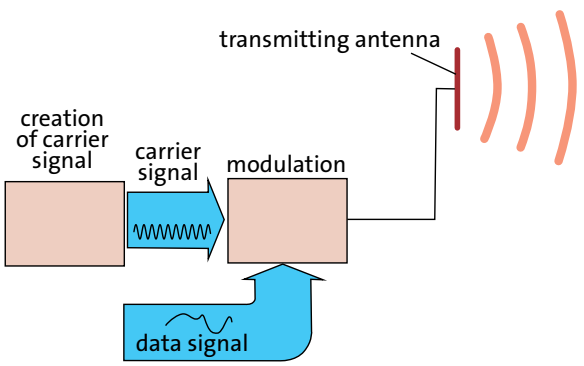

Fig. 6.18 A sine-shaped carrier signal is modulated with the data signal in the data source.

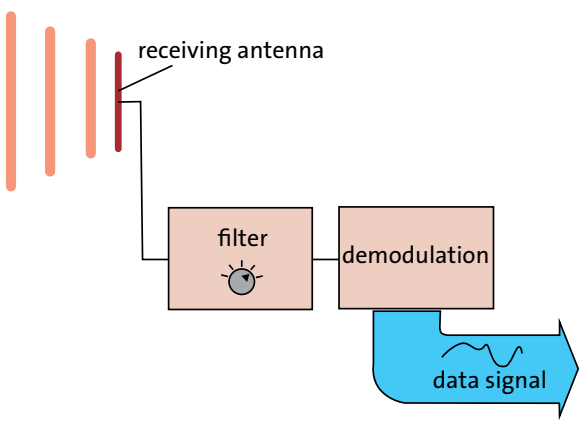

Fig. 6.19 In the data receiver, the frequency range of the desired data source is filtered out of the signal that comes from the antenna. Demodulation occurs afterwards.

To enable several simultaneous data transmissions, every data source is assigned a carrier frequency with a frequency interval (a channel). It may only emit waves with frequencies from this interval.

6.10 Data transmission with electromagnetic waves – modulation

We look at a time function that describes a sine-shaped signal y:

$$y(t) = \hat{y} \cdot \sin(\omega t + \varphi) \, .$$

The function has three „parameters“: the amplitude $\hat{y}$, the frequency f and the starting phase φ. Each of these parameters can be changed over time, i.e. the function can be *modulated* in three ways.

Amplitude modulation

Let's assume we would like to transmit binary data, i.e. an irregular sequence of zeros and ones. The signal is at first available as an electric signal (Fig. 6.20a):

1 → high electric potential
0 → low electric potential

Then, the amplitude of the carrier signal is modulated by setting it to a larger and smaller size in accordance with the initial signal, Fig. 6.20b:

1 → high potential → large amplitude
0 → low potential → small amplitude

An application of the amplitude modulation is the control signal for radio clocks. It comes from a data source that is located close to Frankfurt on the Main. The data source has a reach of approximately 1500 km. The carrier frequency is 77.5 kHz. The amplitude of the wave is now reduced once per second during 0.1 s or during 0.2 s to 25 % of the normal value. This is how the data receiver gets an accurate time signal every second. Why is it sometimes reduced for 0.1 s and sometimes for 0.2 s? Because other data can also be accommodated using this method: the minute, the hour, the date, whether summer or winter time applies, and more.

Amplitude modulation is also used for the old medium- and longwave radio transmission.

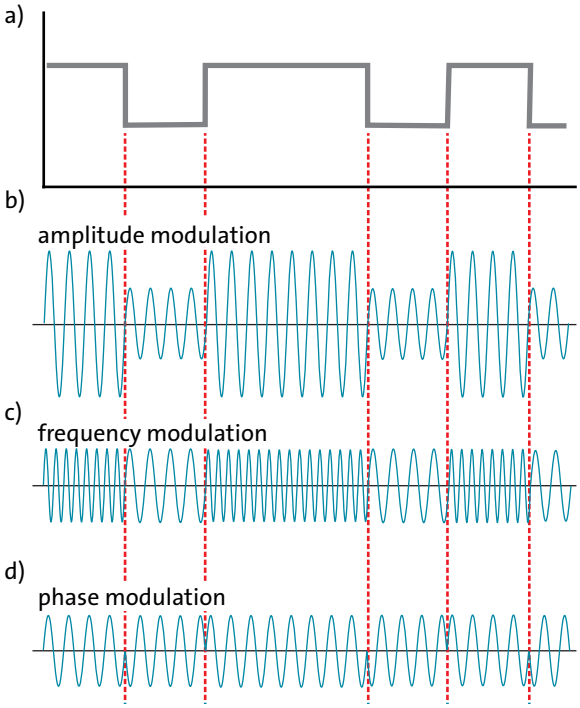

Fig. 6.20 (a) Initial binary signal as a function of time. (b), (c) and (d): amplitude, frequency or phase were changed in accordance with the binary signal. (The modulation signals were not drawn to scale for the sake of clarity. In fact, many more oscillations of the carrier signal are part of each binary sign interval. Also, the relative frequency change during frequency modulation is in reality much lower than in the Figure.)

Frequency modulation

The name is self-explanatory. Here, not the amplitude but the frequency of the carrier wave is changed in accordance with the signal to be transmitted, Fig. 6.20c. Of course, the frequency change has to be kept so small that it remains inside of the assigned frequency interval.

An example for the application of the method is FM radio transmission („FM“ stands for „frequency modulation“). Not binary signs but sound signals are transmitted in this case. The frequency is changed continuously in accordance with the sound signal.

Phase modulation

This is the third possibility to change the carrier wave. Both the amplitude as well as the frequency are left constant whereas the starting phase of the sine-function is changed – again in the accordance with the signal to be transmitted – for example as shown in Fig. 6.20d. During each change from 0 to 1 or from 1 to 0, π is added to the starting phase.

Digital radio transmission, digital television, UMTS

The available channels are scarce and expensive. Therefore, every channel has to be exploited to the maximum, i.e. by transmitting the largest possible data currents. The modulation method has to be as sophisticated as possible.

For this reason, several modulation methods are often combined, for example in case of digital radio transmission (DAB = Digital Audio Broadcasting), for digital television (DVB = Digital Video Broadcasting) or for UMTS (Universal Mobile Telecommunications System).

> Modulation: change of the amplitude, frequency or phase of the carrier wave in accordance with the signal to be transmitted.
>
> The sine components of the modulated wave are within the assigned frequency interval.
>
> The stronger the data current, the larger the required frequency interval.

Exercises

1. The radio clock signal consists of sine-shaped pieces. How many oscillation periods are associated with a short piece (with a duration of 0.1 s) and how many are associated with a long piece (0.9 s)? Which problem would arise if we were to display the radio signal graphically?
2. Search in the Internet the frequency intervals that are assigned to the best-known radio and television transmitters.

6.11 Data transmission with electromagnetic waves – direct and guided waves

Data move from the „producer“ to the „costumer“ in a variety of ways. The possibilities depend above all on the applied carrier frequency. In section 4.14, we learned:

„For the laws of 'geometrical optics' to apply, the diameter of apertures, lenses and mirrors has to be large compared to the wavelength.“

This means: for short waves, there must be an unobstructed view between the transmitting and the receiving antenna. The waves can be bundled by means of parabolic mirrors.

No line of sight is required for long waves. They can only be bundled to a limited extent.

There is a particularity for the light waves, i.e. the shortest waves used: they can be guided over any path by means of a glass fiber.

Data into all directions

To reach many data receivers, which are spread over a large area, simultaneously by means of an antenna, a wave is created that moves away in all horizontal directions. Such a wave becomes increasingly weak with a growing distance from the transmitting antenna. Examples are the antennas of normal radio and television data sources, of mobile telephone systems and UMTS.

Microwave radio relay

Data sometimes have to be transported over a long distance from one place to a single other place. Therefore, the wave is bundled by means of a parabolic antenna. The method is called *Microwave radio relay.* The wavelengths are in the range of a few centimeters. Microwave radio relay antennas are installed on tall buildings or on special antenna towers. Transmitting and receiving antennas have to be located within line of sight of each other.

Data transport through satellites

For a satellite at a distance of approximately 42,000 km from the center of the Earth, the orbiting time is just 24 hours. Hence, its angular velocity is equal to that of the rotation of the Earth around its own axis. If the orbit of such a satellite is in the equator plane and if the rotating direction of the satellite around the center of the Earth is the same as that of the Earth, it is at rest relative to the Earth, Fig. 6.21. It is always located at the same place above the equator and therefore appears motionless from the Earth. Such *geostationary satellites* are ideal data transmission stations. They are

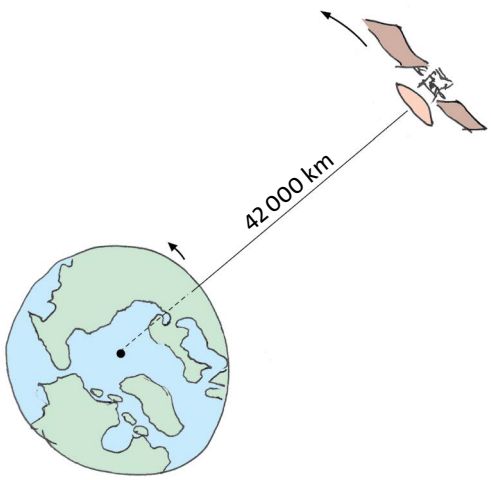

Fig. 6.21 From the perspective of the Earth, a geostationary satellite is standing still.

used for data transmissions over long distances, i.e. for example across continents, but also for the distribution of television programs directly to the television customer.

The data are sent with short waves from a ground station by means of a parabolic antenna towards the satellite. The satellite receives them with its receiving antenna, amplifies the signal and sends it back downwards with its transmitting antenna: either as a strongly bundled beam to another ground station or less strongly bundled to the television customers in an entire country or continent.

The waves that are received this way by means of a parabolic antenna have – similar to directional audio – wavelengths of several centimeters.

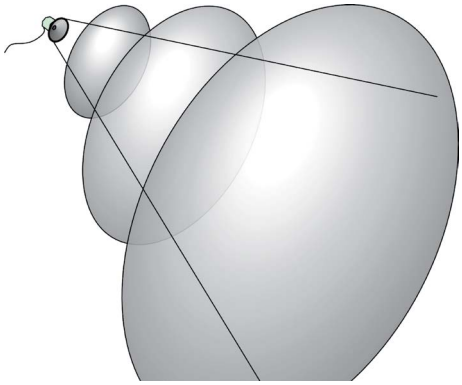

Fig. 6.22 The wave, and consequently its energy, spreads over a constantly increasing area.

Optical waveguides

An optical waveguide is a thin fiber made of a transparent material. The light that is fed in on one end moves through the fiber without escaping laterally, even if the optical waveguide makes curves or loops. (A sound wave behaves similarly in a tube – just try it out.)

Fibers made of quartz glass (SiO_2) with a diameter of around 1/100 mm, i.e. the thickness of a hair, are used for data transmission. Both visible as well as infrared light is used. The light source is a small laser, the data receiver is a photo diode.

Modulation is ensured by switching the light on and off very quickly. Hence, the method is an amplitude modulation.

Compared to electric cables, optical guides are advantageous because they can carry much greater data currents. Data currents of up to 50 Gbits/s can be achieved with one fiber, i.e. approximately 50 times as much as can pass a copper cable. In addition, optical waveguides are less prone to failure than copper cables. Finally, the energy losses are smaller than for copper cables so that fewer amplifiers are needed (see next section).

6.12 Amplifiers

Energy is needed for any type of transport. The truck that transports bricks from the brick factory to the construction site consumes Diesel fuel and hence energy. To make water flow through a water pipe or crude oil through the pipeline, pumps are needed – and such pumps require energy. What happens to the energy that is used for these transports? An energy loss always means that entropy is created somewhere: through mechanical friction, in electric resistors, in chemical reactions.

Energy is needed for the transport of data, too. In most cases, it is supplied by the data source. In other words, the data carry it along as a provision. Therefore, sound waves created by a speaker, the electromagnetic waves that come from an antenna, or the light that comes from the screen of a television carry energy besides the data.

Some data transports are particularly profuse: the wave sent out from the source spreads out over a constantly increasing area, Fig. 6.22.

This applies for example for the sound that comes from a speaker or from a speaking person, or for the electromagnetic waves that are emitted by a television transmitting antenna. Hence, also the energy is spread in the large area. This method is always practical in cases where we seek to reach many data receivers without laying a line to every single one. Most of the energy will of course not arrive at any data receiver but it will be lost.

If the electromagnetic waves are bundled with a parabolic antenna, a larger part of the emitted energy will arrive at the receiving antenna.

At the data receiver always something has to be operated or controlled:

- electric currents have to be induced in a receiving antenna;
- the eardrum of a hearing person has to be set in motion;
- the membrane of a speaker has to be moved.

These processes can only take place if a sufficient amount of energy arrives with the data. If the energy losses in a telephone line are too high or if the radio receiver is too far away from the data source, the arriving energy will no longer be sufficient.

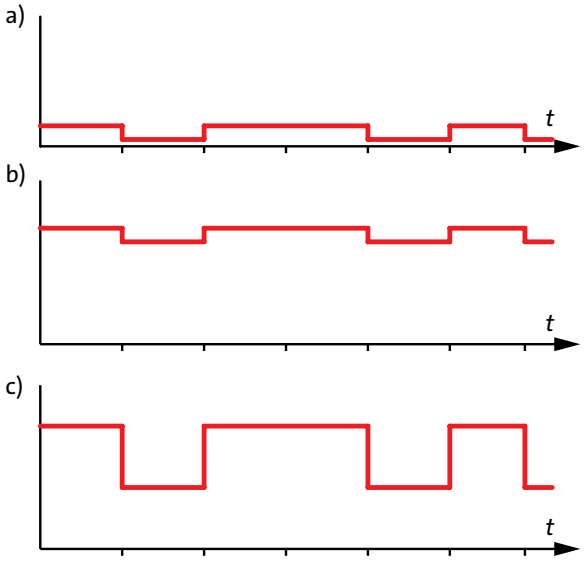

Fig. 6.23 Energy current as a function of time for (a) weak signal (b) weak signal + constant energy current (c) strong signal

Amplifiers are used to ensure that enough energy will arrive at the data receiver in spite of a long distance from the data source. An amplifier has an input and an output for the data. They flow into the amplifier with little energy and leave it with much energy. Hence, the data current receives new provisions.

In an amplifier the energy current that accompanies a data current is increased.

We use the example of the electric amplifier to get a clear idea of how an amplifier works. For the sake of simplicity, we assume the data to be encoded binarily. A „weak signal", which could look like the illustration in Fig. 6.23a, enters into the amplifier.

Here, the energy current is shown as a function of time. The amplifier transforms it into a „strong signal". It is important to understand that the new energy current is not simply *added* to the weak signal as shown in Fig. 6.23b. The result would still be referred to as a weak signal because the differences between the higher and the lower values are still as hard to detect as for the signal in Fig. 6.23a. The amplifier rather has to *multiply* the energy current with a factor that is as large as possible. The result of a multiplication with the factor 6 is shown in Fig. 6.23c.

An amplifier can be characterized by the *amplification factor*, the factor by which the energy current at the output is stronger than at the input.

Fig. 6.24 shows schematically the data and energy current for a CD drive with its amplifier and speaker boxes. The drive supplies an energy current of approximately 0.1 μW. But the speaker boxes require 10 W. Therefore, an amplifier is positioned between the drive and the speaker boxes. The amplification factor of a typical HiFi amplifier is 10^8.

The energy current of the electric signals that come from a radio antenna is typically 1 pW (= 10^{-12} W). The amplification factor for radio reception consequently has to be around 10^{13}.

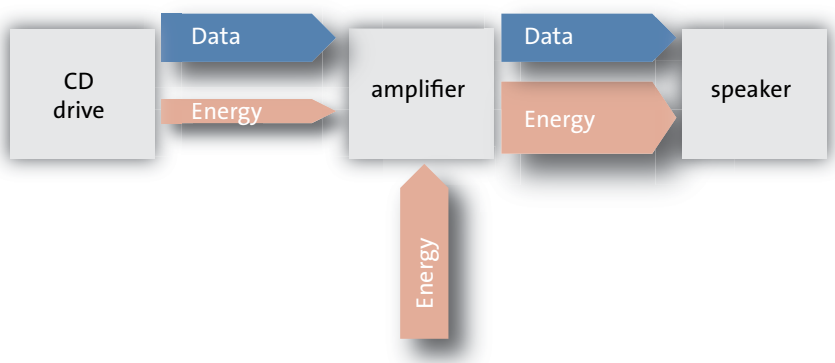

Fig. 6.24 Energy and data current for CD drive, amplifier and speaker boxes