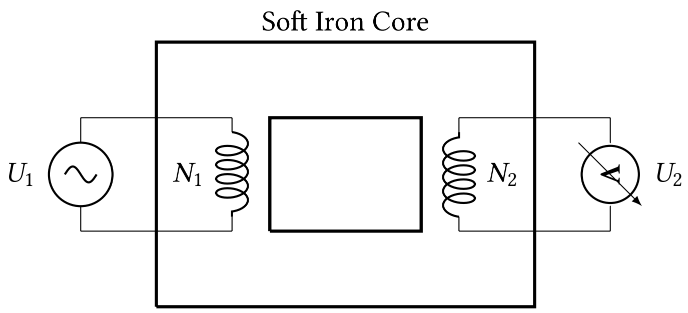
Physics and Engineering

Transformers

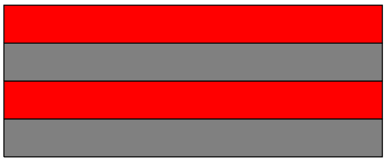
A transformer is a device for changing an alternating voltage from one value to another. It plays a significant role in the distribution and use of electric power for industry and in everyday life. Transformers *step-up* (increase) the voltage produced by an AC generator to very high values to ensure efficient transmission across large distances. Subsequently, substations near to households *step-down* (decrease) the voltage to 230 Volts or other standard values which depend on the particular country. In homes, small transformers in phone or laptop chargers or other electrical devices step-down the voltage further to values that are safe and appropriate for them, often 9 Volts. The underlying principles of such a transformer are those of electromagnetism and electromagnetic induction.

Setup and Working Principle

The basic construction and configuration of a transformer involves two conductive coils of wire wrapped around a closed loop of a soft iron core. One of the two coils is referred to as the *primary coil*, situated on the *primary side* of the transformer. The second coil, on the other end of the soft iron core — the *secondary side*, is called the *secondary coil*. Each coil is characterized by a certain number of windings or turns N_1 and N_2. When the primary coil is connected to a supply of alternating current with a potential difference U_1 across it, an alternating magnetic field is generated around and within the coil. Consequently, the soft iron core is magnetized such that an alternating magnetic flux is present within it. As this changing magnetic flux travels through the closed loop of the soft iron core and reaches the secondary side, it statisfies the requirement for the induction of a potential difference U_2 within the secondary coil. This requirement, as per Faradays' law of induction, is that there either be moving conductor in a permanent magnetic field, or, as is the case here, a stationary conductor in an alternating magnetic field.



The purpose of the soft iron core is to increase the effect of the magnetic field generated as well as to concentrate and guide and the magnetic flux from the primary to the secondary side. Moreover, it reduces magnetic field leakage to a minimum in order to ensure maximum transformer efficiency. However, it should also be mentioned that the conversion and stepping-down or stepping-up of one voltage to another produces not only a magnetic field, but additionally generates thermal energy (heat), which reduces the efficiency of the conversion process. To counteract this, the soft iron core is in fact a sandwich of soft iron plates and insulating material. In such a structure, referred to as *laminated iron*, one layer of soft iron follows one layer of insulation, followed by another layer of soft iron and so on. This ensures maximum possible insulation and efficiency.



The sandwiched layers of the laminated iron core

Transformer Equations

In case of a *step-down* transformer, the voltage U_1 across the primary coil is greater than the induced voltage U_2 across the secondary coil. This can only be the case if the same relationship is true for the turns of the two coils. By contrast, the primary voltage U_1 is less than the secondary potential difference U_2 when the number of turns of the secondary coil N_2 is greater than the number of turns on the primary coil N_1. This form of transformer is then referred to as a *step-down* transformer.

Step-Up Transformer: $U_2 > U_1 \quad N_2 > N_1$

Step-Down Transformer: $U_2 < U_1 \quad N_2 < N_1$

In general, the induced voltage U_2 across the secondary coil can be calculated by slightly altering Faraday's law of induction. This law states that the induced potential difference across a conductor in an alternating magnetic field is equal to the negative of the rate of change of magnetic flux. For transformers, there is an extra factor added to this rate of change: the number of turns N_2 of the secondary coil. This yields the following expression for the secondary voltage U_2:

$$U_2 = -N_2 \cdot \frac{d\Phi_m}{dt}$$

However, the alternating magnetic flux does not only effect the secondary coil, but also causes *self-induction* in the primary coil, as the magnetic flux is also changing on the primary side. Therefore, the above definition for the induced voltage U_2 on the secondary side can also be used to determine the self-induced voltage U_{ind} across the primary coil. This induced voltage is then equal to the negative of the voltage U_1 across the primary coil:

$$U_{ind} = -N_1 \cdot \frac{d\Phi_m}{dt} = -U_1 \qquad \Rightarrow \qquad \frac{d\Phi_m}{dt} = \frac{U_1}{N_1}$$

The right side of this equation can then be substituted for the rate of change of magnetic flux in the expression given above for the secondary voltage U_2:

$$U_2 = -N_2 \cdot \frac{d\Phi_m}{dt} = -N_2 \cdot \frac{U_1}{N_1}$$

This finally yields the following relationship, referred to as the *first transformer law*, where the minus sign indicates a 180° or π radians phase shift between the two voltages U_1 and U_2, as a result of them being wound in the same direction:

$$\frac{U_1}{N_1} = -\frac{U_2}{N_1}$$

Were now a load to be connected on the secondary side (e.g. a lamp), the induced voltage U_2 would cause an alternating current I_2 to start flowing through the secondary coil. As a result, there would also be an alternating magnetic field generated on the secondary side. This magnetic field would in turn also magnetize the soft iron core, leading to a very complex phase relationship between the magnetic flux created by the primary coil and the magnetic flux created by the scoendary coil. In case of an ideal transformer with 100% percent efficiency and no loss energy of electrical energy to thermal energy, the total power of the primary coil P_1 equals the power of the secondary coil P_2. Given that the electrical power of a circuit is equal to the voltage across it multiplied by the current flowing through it, one can determine the following relationship — known as the *second transformer law* — between the voltage and current of the two sides of a transformer:

$$P_1 = P_2 \qquad \Rightarrow \qquad U_1 \cdot I_1 = U_2 \cdot I_2 \qquad \Rightarrow \qquad \frac{U_1}{U_2} = \frac{I_2}{I_1}$$

One last thing to mention is that in reality, no transformer is really ideal, thus the effective voltage U_{eff} across the primary coil is actually equal to the maximum possible primary voltage U_{max}, divided by the square root of two (this can also be used to calculate the maximum voltage given the effective voltage):

$$U_{eff} = \frac{U_{max}}{\sqrt{2}}$$

Power Transmission across the Country

Unfortunately, no conductor is ideal. There is always some portion of electrical power generated by power stations that is lost during transmission of the power across the country. Especially over great distances the loss in power can be very significant, such that the efficiency is too low for practical use. Thus, great care must be taken to ensure maximum efficiency. One way of doing so is to increase the voltage during transmission. To see why, it must first be discussed how the power P_{lost} can be calculated. In general, electrical power is defined as the voltage across a conductor multiplied by the current flowing through it. Moreover, voltage may be defined not only as a potential difference, but also as the product of current and resistance. This leads to the following equation for the power lost during transmission across the country:

$$P_{lost} = U \cdot I = R \cdot I \cdot I = R \cdot I^2$$

The efficiency of transmission may then be calculated as the ratio between the power lost P_{lost} and the power P generated by a given power station (this value could be multiplied by 100 percent to get a direct percentage value for the ratio between the power lost and the power generated):

$$\frac{P_{lost}}{P}$$

When expanding both variables, it can be found that the efficiency is equal to the power generated, multiplied by the resistance and divided by the square of the voltage. This shows that it is not necessary to explicitly know the power lost. It can be calculated already from the value of the power generated.

$$\frac{P_{lost}}{P} = \frac{R \cdot I^2}{U \cdot I} = \frac{R \cdot I}{U} = \frac{R \cdot I \cdot U}{U \cdot U} = \frac{P \cdot R}{U^2}$$

What this shows is that the power-lost-to-power-generated ratio is inversely proportional to the voltage generated. A greater voltage will thus cause less loss in power. Proof for 20 kilo-volts, 100 kilo-volts and 380 kilo-volts at 1 giga-watts of power transmitted through a conductor with 50 Ω of resistance are given below.

20 kV: $\frac{P \cdot R}{U^2} = \frac{1 \cdot 10^9 \cdot 50}{(20 \cdot 10^3)^2} \cdot 100\% = 12500\%$

Conclusion: The power lost is 12500 % of the power generated (none will reach the end user).

100 kV: $\dfrac{P \cdot R}{U^2} = \dfrac{1 \cdot 10^9 \cdot 50}{(100 \cdot 10^3)^2} \cdot 100\% = 500\%$

Conclusion: Still none.

380 kv: $\dfrac{P \cdot R}{U^2} = \dfrac{1 \cdot 10^9 \cdot 50}{(380 \cdot 10^3)^2} \cdot 100\% = 35\%$

Conclusion: Only 35 % of the power generated is lost, such that it will reach the end user with an efficiency rating of 65 %.

Broadcasting

Broadcasting is the transmission of information — mostly radio and television signals — via *radio waves*. Radio waves are found on the lower end of the electromagnetic spectrum, in a frequency range of about $3 \cdot 10^4$ to $3 \cdot 10^9$ Hz. There are three categories into which radio waves are generally divided: *ground* or *surface waves*, *sky waves* and *space waves*.

Ground Waves

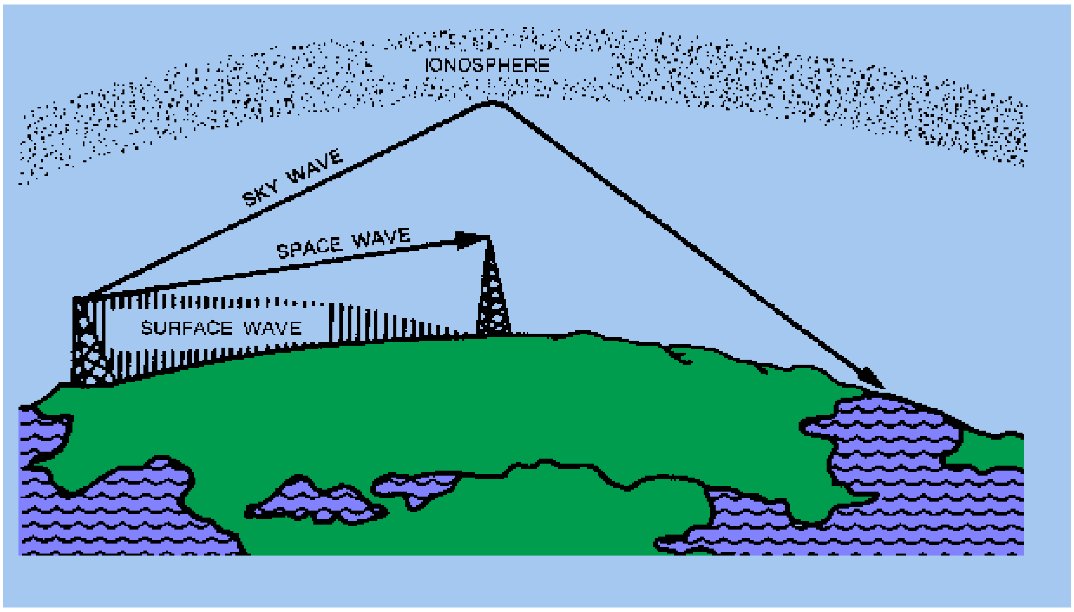
Ground waves (also referred to as surface waves) have the longest wavelength and lowest frequency of the three categories of radio waves, with their frequency typically found at around or below 3 mega-hertz ($3 \cdot 10^6$ Hz). Ground waves are most commonly used for long-distance communication and broadcasting. The reason has to do with the diffraction of waves — the phenomenon whereby waves are bent when passing through gaps or around obstacles. The degree to which a wave is diffracted depends on its wavelength λ. In this case, a greater wavelength causes a greater degree of diffraction to occur. As ground waves have the highest wavelength of all radio waves, they are diffracted most as they travel on the surface of the earth, following the curvature of the surface and continuously being bent.

Sky Waves

Sky waves have a higher frequency than ground waves, typically in a range of 3 to 30 mega-hertz ($3 \cdot 10^6 - 3 \cdot 10^7$ Hz), and thus have a shorter wavelength. This also means that they are less suitable for long-distance communication taking place near to the surface of the earth, as a shorter wavelength would cause such a sky wave to experience less diffraction. However, at this frequency the waves do have the ability to propgate through the earth's atmosphere, up to a layer known as the *ionosphere*. The ionosphere is an electrically charged region above the atmosphere, around 50 to 500 kilometers above the surface of the earth, where ultraviolet radiation from the sun results in a high number of positively or negatively charged particles (*ions*) to be found at such altitudes. The intensity of solar radiation and thus ionization changes over time, as the level of radiation from the sun is, of course, less during the night and greater during the day. When sky waves reach the ionosphere, they are *reflected* similar to the way in which light is reflected in the process of total internal reflection. After reflection from the ionosphere, sky waves propagate back to the earth's surface, either to reach their destination directly (*single-hop-transmission*) or to be reflected once more towards the ionosphere after being received by an antenna on the surface of the earth and subsequently re-transmitted.

Space Waves

Space waves have the highest frequency of all radio waves and thus experience practically no diffraction at all. Therefore, they cannot be used for long-distance near-surface transmission as ground waves can be. Moreover, they are no longer reflected by the ionosphere, but can now pass through it. As a result, space waves are commonly used for satellite communication or direct transmission of signals betwee antennae and in general any form of *line-of-sight-communication*, where the receiving antenna is within the transmitter's line of sight (i.e. a direct, straight path of tranmission) and where diffraction may even be unwanted. This is also a reason why short-wavelength radio waves similar in frequency to space waves are used for the *Bluetooth* protocol between smartphones or other electronic devices such as printers, loudspeakers or computer keyboards. The reason why Bluetooth has only a very short range of usability should now be clear.



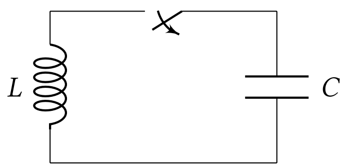
The three categories of radio waves

LC-Circuits

Electromagnetic waves are transverse waves that travel at the speed of light and consist of a magnetic field and an electric field component, whose field strengths oscillate perpendicular to each other and to the direction of the wave's motion. They find a plethora of applications in everyday life and especially in the transmission of data or signals for television, radio or telecommunications. The most common method of generating such electromagnetic waves is what is known as an *LC-circuit*.

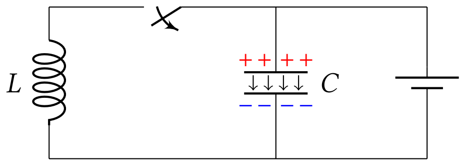
Setup of an LC-Circuit

The configuration or setup of an LC-circuit consists of two main components: a *capacitor* and an *inductor* (a coil). An important property associated with the inductor is its *self-inductance* L, measured in henries $[H]$. On the other hand, the capacitor — an electrical component used to store a certain electrical charge — is characterized by its capacitance C, measured in farads $[F]$. The capacitor and the inductor are connected in a closed series circuit, where also a switch may be present to interrupt the flow of current.



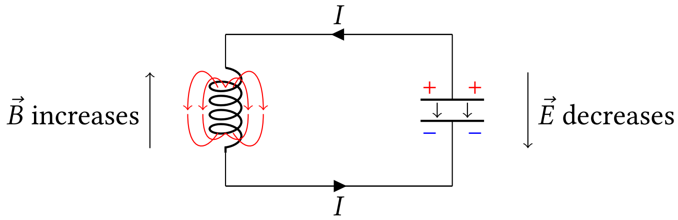
0. Charging the Capacitor

Initially, to eventually produce electromagnetic waves, the capacitor must be given a certain electrical charge. For this, the switch is opened and the capacitor is connected to a power supply, which charges the plates of the capacitor. As it does so, one of the plates of the capacitor is charged negatively while the other plate receives a positive charge. Consequently, a homogenous electric field is produced between the positive and the negative plate of the capacitor. The property of homogeneousness refers to the fact that the electric field is equal in strength and density at every point between the two plates.



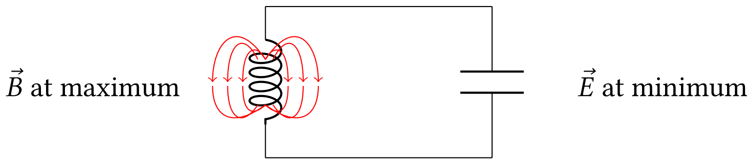
1. Closing the Switch

Next, the battery is removed and the switch is closed. Current then starts to flow and the capacitor gradually discharges. As current flows through the turns of the inductor, a magnetic field is produced within and around it. As this inductor is essentially a coil or solenoid, the shape of the magnetic field generated is similar to that of a bar magnet. At this stage, the electric field strength $\vec{E}$ decreases with time while, simultaneously, the strength of the magnetic field $\vec{B}$ generated by the current flowing through the inductor is continuously increasing.



2. Capacitor discharges

As the capacitor gradually discharges and is almost entirely void of charges, the magnetic field strength $\vec{B}$ also nears its maximum strength. At one point, the capacitor is then *fully* discharged, such that the current eventually stops flowing. At this point, the magnetic field reaches its maximum and then begins to decrease in strength.

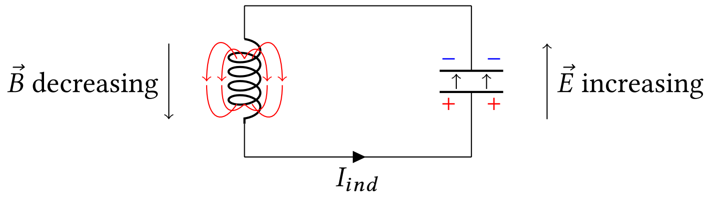


3. Induction of Current

This decrease in the magnetic field strength signifies a change in the magnetic flux in the coil, leading it to *self-induce* a voltage and consequently a current. The direction of the induced current I_{ind} is the same as that of the original current I. The reason for this can be derived from Lenz' law, which states that:

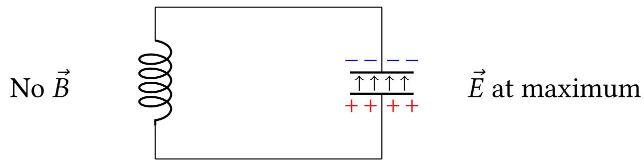
> The direction of the induced voltage is such that, were an induced current able to flow, it would oppose the change that caused it.

The change that caused the induction of the the current was the *breaking-down* of the magnetic field of the inductor. As stated by Lenz law, the induced current always flows in the direction that acts against the change that caused it. It thus must try to keep the magnetic field constant and stop its degeneration. The only way for it to do so is to keep flowing in the same direction, as this is the direction that caused the field to increase initially. The side-effect of this is the most important property of the LC-circuit: the capacitor is gradually re-charged in the opposite way — with opposite polarity.



4. Capacitor fully re-charged

At last, the capacitor is again fully charged, such that the electric field $\vec{E}$ is again at its maximum, while the magnetic field $\vec{B}$ is no longer present at all, such that $|\vec{B}| = 0$.



The whole cycle would repeat once more, just that the polarity of the capacitor is now reversed. After the next half-cycle, the polarity would again match its original configuration. If energy losses are neglected, this oscillatory behaviour could go on ad infinitum. In conclusion, it can be said that, for an LC-circuit …

- …the charges oscillate between the plates of the capacitor
- …the energy oscillates between the electric field $\vec{E}$ and the magnetic field $\vec{B}$
- …the field strengths of both the electric field $\vec{E}$ and the magnetic field $\vec{B}$ oscillate

The frequency with which this oscillation takes place can be determined using an equation referred to as the *Thomson formula*. Its derivation is fairly simple, if it is known that the angular frequency ω of the charge and energy oscillation is equal to the inverse square root of the product of the self-inductance L of the coil and the capacitance C of the capacitor. Moreover, it must be remembered that the angular frequency ω of any object or system undergoing periodic behaviour is equal to two π times the frequency f. Thus, the frequency f must be equal to ω divided by 2π.

$$\omega = \frac{1}{\sqrt{L \cdot C}} \qquad \omega = 2\pi \cdot f \Rightarrow f = \frac{\omega}{2\pi}$$

Inserting the angular frequency of an LC-circuit as defined above into the right equation yields the Thomson formula for the determination of the frequency of oscillation of an LC-circuit:

$$f = \frac{1}{2\pi} \cdot \frac{1}{\sqrt{L \cdot C}}$$

The Hertz Dipole

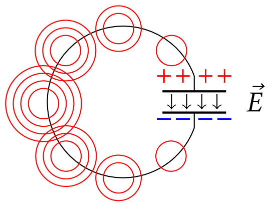
It was just shown how an LC-circuit can be used to generate a charge and energy oscillation between a capacitor and an inductor. Moreover, it was stated that the angular frequency ω of the oscillation is equal to the inverse square root of the self-inductance L multiplied by the capacitance C. Taking a closer look at this definition, it can be seen that to increase the angular frequency, one would either have to decrease the self-inductance L of the coil or decrease the capacitance C of the capacitor, or both.

Decreasing L

To decrease the self-inductance L, one needs to reduce the number of turns of the coil. The minimum number of turns is, of course, one turn. Thus, the inductor is essentially turned into a single loop of wire. Consequently, the field lines of the magnetic field produced are structured in concentric circles around the wire, such that the magnetic field strength B decrases with an increasing distance r from the wire, as per the defintion of the field strength B of the magnetic field generated around a single wire:

$$B = \frac{\mu_0}{2\pi} \cdot \frac{I}{r}$$

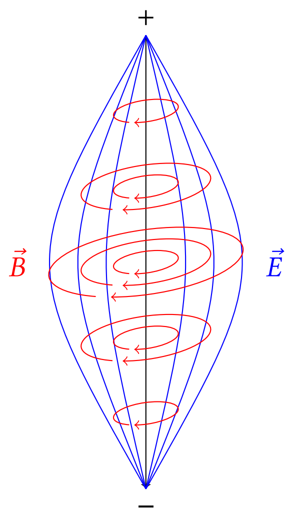
Furthermore, it can be observed that the magnetic field is initially very small closer to the positive plate, reaches its maximum in the middle of the wire and finally decrases again as it nears the negative plate (the same could be said for electrons moving in the opposite direction). This fluctuation in magnetic field strength is due to the current being small near the plates and greatest at the midpoint.



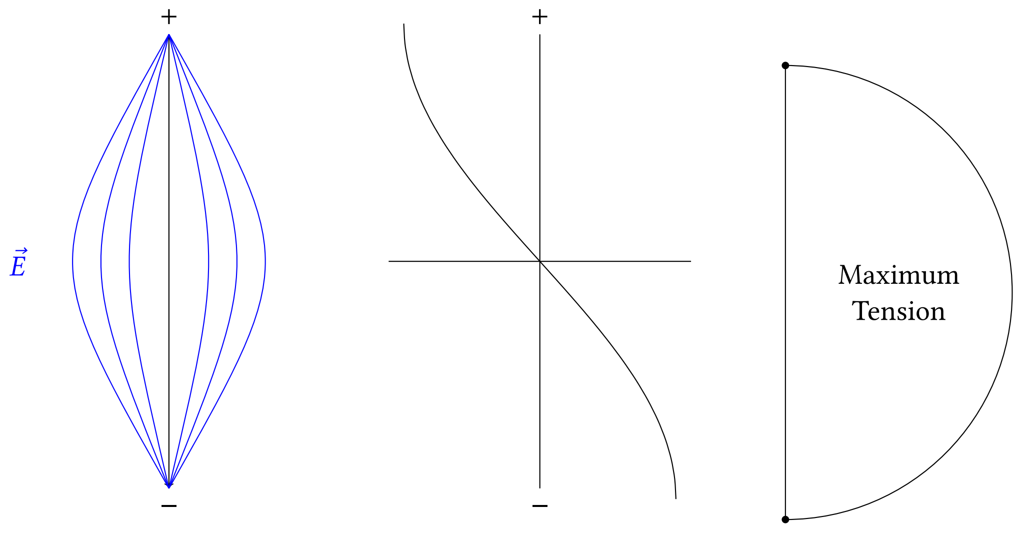
Decreasing C

The capacitance C is determined by the distance between the two plates of the capacitor, as well as the size of the plates. Thus, one would first decrease the size of the plates to two single points of the same cross-sectional area as that of the wire. Then, one would entirely open up the single loop of wire and thus LC-circuit, to increase the distance to a maximum. As one does so slowly, one would notice that the electric field between the plates of the capacitor begins to spread out into space. When the loop is then opened up to an entirely straight wire, the LC-circuit can be described by the following properties:

- It is a **straight** wire, but still essentially an inductor coil (with onely one loop, in a very deformed shape)
- The plates of the capacitor have been reduced to **single points**
- The electric field and magnetic field are **no longer separated**. Rather, they **overlap**.
- Both the electric and the magnetic field go **around the wire**. This was not the case for the electric field before; it used to only be present between the plates of the capacitor.
- The **length** of the wire remains the only variable to change the frequency, as the capacitance depends on the distance between the plates.

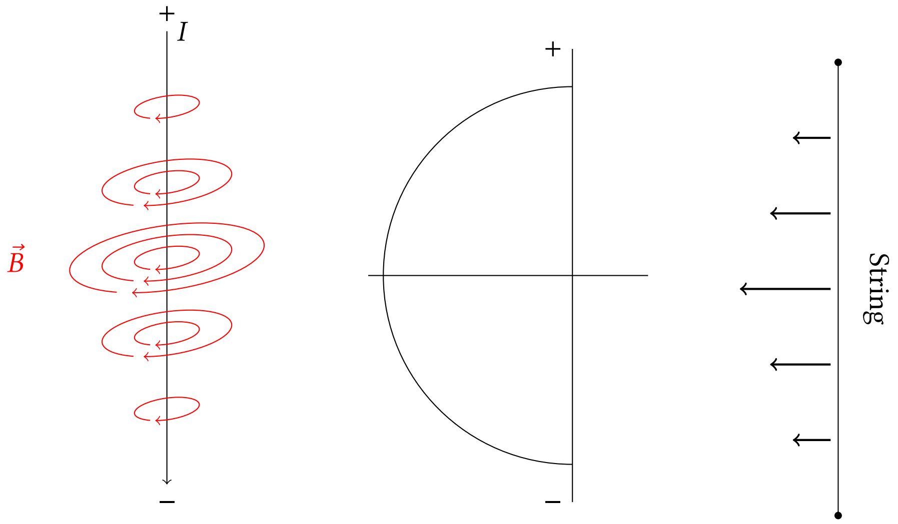


There are two separate scenarios of the Hertz Dipole's cycle that will be observed in further detail now. The first is the case in which all negatively charged particles are concentrated at one plate — then the negative pole — before any current starts to flow. At this stage, the antenna has the maximum positive potential difference at the positive end, where fewer electrons are, while it has its maximum negative potential difference at the negative plate. The potential difference decreases from both ends towards the midpoint, where the voltage is then zero. As the charge flow and concentration reverses during the Hertz Dipole's cycle of oscillation, the polarity of the voltage reverses as well. Thus, the potential difference distribution across the antenna can be seen as a standing wave with two nodes at the + and − plates, that is 90° out of phase, or with two antinodes at the poles and one node at the midpoint. At this stage of no current flowing and maximum potential difference (as all charges are at one plate), comparisons can be made with a violin string that is stretched to a maximum, such that it is experiencing the greatest amount of tension and is storing the greatest amount of energy as elastic potential energy, similar to how the current has the maximum potential to flow at this point.



The next stage of the open LC-circuit's oscillatory cycle is the situation in which current is flowing through it. The distribution of current again looks like a standing wave with two current nodes (+ and −) at the ends of the antenna, such that it is 90° out of phase with the voltage standing wave. The current as well as the magnetic field generated is greatest at the center of the dipole, while it decreases in strength nearer towards either plate. To again use the analogy of a violin string, one could compare this situation with the velocity distribution of a violin string after it is let go from the previous position of maximum tension given before. The string has the greatest velocity in the center.

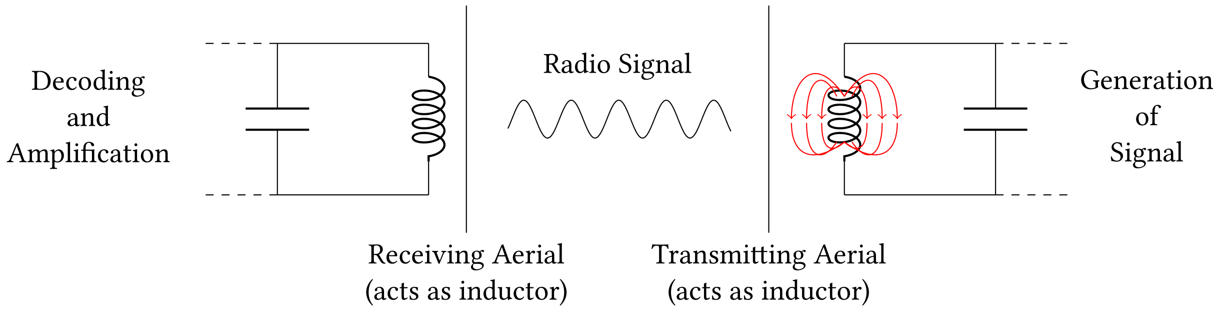
As the electrons flow towards the opposite plate, the polarity is eventually reversed. Then, the voltage and current standing wave as well as the violin standing waves will be opposite in direction. In general, the similarity between the distribution of current strength of a Hertz Dipole and the shape of a standing wave with two nodes (+ and −) earns the antenna another name: *lambda half dipole*.



Transmission of Electromagnetic Waves

LC circuits, in their closed and open forms, are at the basis of all broadcasting and transmission of electromagnetic waves. For any form of transmission, there must be a receiving antenna and a transmitting antenna. To begin the creation and transmission of a radio signal, one closed LC circuit is placed in the proximity of Hertz Dipole antenna. The alternating magnetic field of the closed circuit's inductor resulting from its the charge and energy oscillation causes high-frequency alternating current to be induced in the transmitting aerial. This is due to the fact that the aerial is essentially an inductor coil with only one winding, bent to a straight line. It is the alternating magnetic flux from the inductor of the closed LC circuit that causes the induction of a current in the aerial. This current, in turn, results in a changing magnetic and electric field around the aerial. If the length of the antenna is comparable to the wavelength λ of the electromagnetic wave, the major part of the antenna's electric and magnetic field energy is sent off as an electromagnetic wave. It should also be mentioned that the wavelength λ in this case depends on the frequency of the charge oscillation within the Hertz dipole, as it is the frequency of this oscillation that determines the frequency f of the electromagnetic wave. The relationship between the wavelength and the frequency then stems from the fact that the velocity c of the wave (speed of light) is equal to the product of λ and f: $c = \lambda \cdot f$.

The *receiving arial* intercepts the electromagnetic waves transmitted from the transmitting antenna. It can receive the signal as the alternating magnetic field strength of the electromagnetic wave will induce high frequency alternating current in the receiver and thus cause high frequency charge oscillation to occur between the conductor plates of the antenna, resulting in a time-varying electric and magnetic field strength around the aerial. The information is then interpreted, decoded and processed further (e.g. amplification) by a second LC circuit. Lastly, it remains to say that the frequency at which the receiver intercepts electromagnetic waves is determined by varying the capacitance of the capacitor. Thus, when *tuning* your radio, you are changing its capacitance and thereby adjusting the radio receiver to receive the desired radio signal frequency that a particular radio station uses.



Transmission of Information

It was just explained how electromagnetic waves could be transmitted by means of an antenna, also known as a Hertz Dipole or simply an open LC-circuit. However, it was not mentioned how *information* or *intelligence* could be transmitted. The two main methods of altering a signal to incorporate some form of information are Amplitude Modulation (AM) and Frequency Modulation (FM).

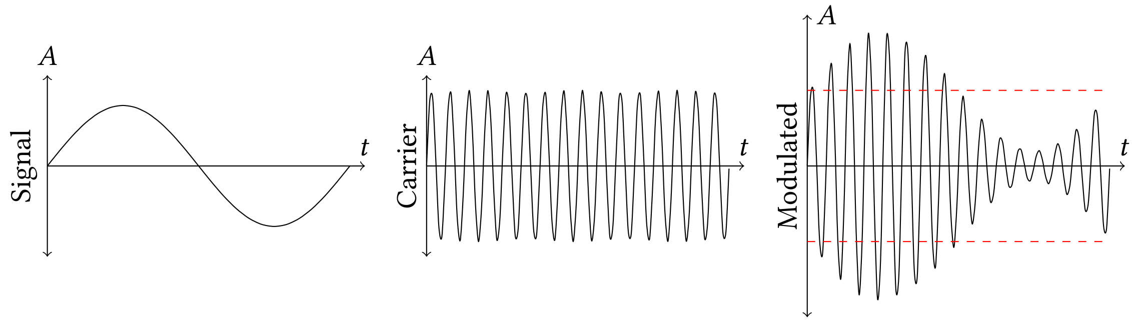
Both methods of modulation modulate a high-frequency *carrier wave* to transport a certain *low-frequency* signal, such as that of human speech, in a particular way from a transmitting antenna to a receiving antenna. Also, beforehand, it should be defined what the *bandwidth* of any modulated signal is. For this it must be known that frequency modulation and amplitude modulation cause *sidebands*, additional spectral components, to be added to the signal as a by-product of their modulation. Sidebands are also referred to as *harmonics* or *overtones* when their frequencies are in harmonic ratios to the *fundamental frequency*, i.e. when their frequencies are integer multiples of the frequency of the carrier wave. The bandwidth of a signal is subsequently a measure of the range or the amount of property a wave and all of its additional spectral components occupy or *consume* on the frequency spectrum. More precisely, it is defined as the difference between the highest and lowest frequency present in the signal.

Amplitude Modulation

Amplitude Modulation (AM) is a method of transmitting information via radio waves that was first succesfully implemented at the end of the 19th century.For amplitude modulation, the *amplitude* of a carrier wave with a pre-defined, fixed frequency is modulated (changed) according to the signal to be transmitted. The degree of modulation of the carrier wave over time is exactly *equal* to the deviation in the amplitude of the signal. Therefore, were the signal wave to reach its peak at a certain point in time, one would find that the difference between the carrier wave's unmodulated amplitude and its actual amplitude at that instant would be exactly equal to the value of the signal wave. Similarly, the deviation in the amplitude of the carrier wave is negative, i.e. towards the zero line, when the signal falls to a value below zero. An important property of amplitude modulation is that while the amplitude of the carrier is altered continuously, its frequency and phase remain *unchanged*. Lastly, it should be mentioned that to finally decode an amplitude-modulated signal at the receiving end of a transmission process, an analog filter or *de-modulating circuit* filters out the high-frequency components of the carrier wave, leaving only the lower-frequency signal to be interpreted.

A basic definition of amplitude modulation as a function of time for a carrier wave with amplitude A_c and angular frequency ω_c as well as a periodic modulator signal with amplitude A_m and angular frequency ω_m is given below. ω is equal to $2\pi f$, where f is the frequency of any signal or wave.

$$f(t) = (A_c + A_m \cdot \sin(\omega_m \cdot t)) \cdot \sin(\omega_c \cdot t)$$



The frequency range of amplitude-modulated radio waves is roughly 540 to 1700 kHz ($5.4 \cdot 10^5 - 1.7 \cdot 10^6$ Hz). AM waves are found in a lower frequency range than FM waves, thus they also have a longer wavelength. A longer wavelength is a favorable property for long-range transmission, as the degree to which a wave is *diffracted* and can thus follow the curvature of the earth increases with the length of the wavelength (and thus decreases with an increasing frequency). In comparision, frequency modulated radio waves have a more limited area or range of coverage. The bandwidth requirement of an AM signal is generally defined as two times the highest modulating frequency. In AM radio broadcasting, the modulating signal has a bandwidth of 15 kHz, thus the bandwidth of an amplitude-modulated radio signal is 30 kHz. The bandwidth consumption of an FM signal is comparably much higher (around 80 kHz), thus there can be more AM stations than FM stations within a certain frequency range.

The greatest disadvantage of amplitude modulation is that AM signals are highly susceptible to *noise* and *static*. The reason for this is that AM signals store information in the amplitude of the carrier wave, which can be affected by electromagnetic interference and disturbances. For example, a lightning strike may result in unwanted changes to the amplitude and can distort the signal. This does not happen with frequency modulation, which is why FM has become the more popular of the two forms of modulation for radio and generally audio signals. An interesting fact to mention here is that television signals are actually transmitted by both amplitude modulation and frequency modulation. The audio component of the signal is transmitted via a frequency modulated carrier, while the video component is carried by an amplitude modulated wave. This is why, often, during a thunderstorm, the video signal of your television may be distorted while the audio signal remains perfectly intact.

Some last aspects of amplitude modulation that may be of interest are cost and power consumption. Amplitude modulation was invented already at the end of the 19th century, thus the circuitry required is much less complicated. This also reduced the cost of implementation. Thus, many cheap radios will use AM receivers as they are the more economic option compared to FM.

In summary:

- **Advantages of Amplitude Modulation:**
  - **Greater range of coverage** due to a greater wavelength and greater diffraction around the surface of the earth.
  - **Cheap and simple** to implement.
  - **Lower bandwidth** means more available channels of communication compared to FM.
- **Disadvantages of Amplitude Modulation:**
  - **Highly susceptible to noise**, resulting in a greater degree of distortion from electromagnetic interfernce and static, causing an overall much lower quality of the signal transmitted.

Frequency Modulation

Frequency modulation is the most popular and most commonly encountered form of signal modulation in radio transmission and broadcasting. While amplitude modulation varies the *amplitude* of a high-frequency carrier wave according to the low-frequency signal information to be transmitted, frequency modulation changes the *frequency* of the carrier wave in accordance with the signal. When the signal is above zero, the frequency of the carrier is increased in direct proportion to the signal. When the signal takes on a negative value, the frequency of the carrier wave is reduced by exactly the value of the signal. As such, were a carrier wave with a frequency f_c to be frequency-modulated by a periodic signal with an amplitude A_m, then the frequency of the carrier would deviate between a maximum of $f_c + A_m$ Hertz and a minimum of $f_c - A_m$ Hertz. In this case, while the frequency of the carrier wave is constantly changing, its amplitude and phase remain constant.

The definition of a frequency-modulated signal $f(t)$ as a function of time is given below, where A_c refers to the amplitude of the carrier — which stays constant — while A_m denotes the amplitude of the modulator wave, i.e. the signal — assuming it is perfectly periodic. In that case, the amplitude of the signal A_m also determines the maximum frequency deviation of the carrier wave Δf_c, measured in Hertz. In the below equation, ω_c is the angular frequency of the carrier wave and ω_m that of the signal.

$$f(t) = A_c \cdot \sin((\omega_c + A_m \cdot \sin(\omega_m \cdot t)) \cdot t)$$

Frequency modulation is performed on carrier waves in a frequency range of 88 to 108 MHz ($8.8 \cdot 10^7 - 1.08 \cdot 10^8$ Hz). As such, their operating frequency is much higher than that of amplitude-modulated signals. This also means that they are less suitable for long-range transmission and have a generally lower area of coverage compared to AM signals, as their higher frequency indicates a shorter wavelength and thus a lesser degree of diffraction and ability to follow the curvature of the earth. The bandwidth B of an FM signal is formally defined as two times the sum of the highest modulating frequency f_m and the maximum frequency deviation Δf_c of the carrier wave:

$$B = 2 \cdot (f_m + \Delta f_c)$$

The typical bandwidth of an FM signal is 80 kHz. This confirms the previous statement given that one could fit more 30 kHz AM channels into the same range of frequencies than one could fit 80 khz FM stations. Moreover, one benefit derived from a higher bandwidth is a greater signal quality, as more sidebands can be transmitted alongside the carrier frequency.

As frequency modulation stores the information of the signal in the frequency (deviation) of the carrier wave, the intelligence transmitted is unaffected by changes to the carrier's amplitude. Static, noise or electromagnetic interference from sources like lighting are thus no longer an issue. As a result, signal quality is drastically higher for frequency modulated signals compared to their amplitude-modulated counterparts, also due to the fact that FM signals have a greater bandwidth at their disposal. However, FM circuitry requires more power, is more expensive and a lot more complex than AM transmitters and receivers. This may play a role under certain circumstances.

In summary:

- **Advantages of Frequency Modulation:**
  - **Resistance against noise**, static and other forms of electromagnetic interference such as lighting. Increases signal quality.
  - **Greater bandwidth** also increases signal quality.

- **Disadvantages of Frequency Modulation:**
  - **Lower range of coverage** as the frequency range of carrier waves is higher for frequency modulation than it is for amplitude modulation. Higher frequency means shorter wavelength, which in turn means less diffraction on the surface of the earth or around mountains.
  - **Expensive and complex**.
  - **Greater bandwidth** not only means better signal quality, but also fewer channels available, which may be a detriment.