1

Introduction

1.1 A Brief History of Antennas

Work on antennas started many years ago. The first well-known satisfactory antenna experiment was conducted by the German physicist Heinrich Rudolf Hertz (1857–1894), pictured in Figure 1.1. The SI (International Standard) frequency unit, Hertz, is named after him. In 1888, he built a system, as shown in Figure 1.2, to produce and detect radio waves. The original intention of his experiment was to demonstrate the existence of electromagnetic radiation. In the transmitter, a variable voltage source was connected to a dipole (a pair of 1 m wires) with two conducting balls (capacity spheres) at the ends.

The gap between the balls could be adjusted for circuit resonance as well as for the generation of sparks. When the voltage was increased to a certain value, a spark or break-down discharge was produced. The receiver was a simple loop with two identical conducting balls. The gap between the balls was carefully tuned to receive the spark effectively. He placed the apparatus in a darkened box to see the spark clearly. In his experiment, when a spark was generated at the transmitter, he also observed a spark at the receiver gap at almost the same time. This proved that the information from location A (the transmitter) was transmitted to location B (the receiver) in a wireless manner – electromagnetic (EM) waves! The information in his experiment was actually in binary digital form by tuning the spark on and off. This could be considered as the very first digital wireless system that consisted of two of the best-known antennas: the dipole and the loop. For this reason, the dipole antenna is also called Hertz (dipole) antenna (Figure 1.2).

While Heinrich Hertz conducted his experiments in a laboratory and did not quite know what radio waves might be used for in practice, Guglielmo Marconi (1874–1937, pictured in

Antennas: From Theory to Practice, Second Edition. Yi Huang.

Companion website: www.wiley.com/go/huang_antennas2e

Figure 1.1 Heinrich Rudolf Hertz

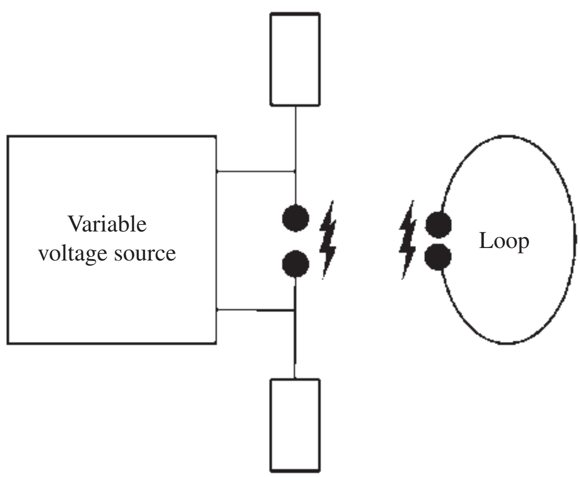

Figure 1.2 1887 experimental setup of Hertz's apparatus

Figure 1.3), an Italian inventor, was the man who developed and commercialized wireless technology by introducing a radiotelegraph system, which served as the foundation for the establishment of numerous affiliated companies worldwide. His most famous experiment was the transatlantic transmission from Poldhu, UK, to St Johns, Newfoundland, in the USA in 1901 employing un-tuned systems. He shared the 1909 Nobel Prize in Physics with Karl Ferdinand Braun 'in recognition of their contributions to the development of wireless

Figure 1.3 Guglielmo Marconi. Source: https://commons.wikimedia.org/wiki/File:Marconi_1909.jpg#/media/File:Marconi_1909.jpg

telegraphy'. Monopole antennas (near quarter wavelength) were widely used in his experiments, thus vertical monopole antennas are also called Marconi antennas.

During World War II, battles were won by the side that was first to spot enemy airplanes, ships, or submarines. To give the Allies an edge, British and American scientists developed radar technology to 'see' targets from hundreds of miles away, even at night. The research resulted in the rapid development of high-frequency radar antennas, which are no longer just wire-type antennas. Some aperture-type antennas such as reflector and horn antennas were developed, an example is shown in Figure 1.4.

Broadband, circularly polarized antennas, as well as many other types, were subsequently developed for various applications. Since an antenna is an essential device for any radio broadcasting, communication, and radar systems, there has always been a requirement for better or new antennas to meet existing and emerging applications.

For example, the cellular radio communication system is moving to its 5th generation (5G), the operational frequencies are extended from sub-6 GHz (e.g. 698–960, 1710–2690, and 3300–3800 MHz) to millimeter waves. The number of antennas in both the mobile portable and the base station is increased significantly to form massive multiple input and multiple output (MIMO) system to dramatically increase the communication data rate and capacity. This means massive new challenges to antenna designers: the antennas are to be placed in a relatively small device, such as a smartphone, and need to perform well at different frequencies (including 3G and 4G mobile frequencies) at the presence of other electronic systems (e.g. Wi-Fi, GPS, cameras, and a large display) and human body/hands. At millimeter waves, the antennas are also expected to produce beaming forming and steering functionalities to combat increased path loss which poses one of the main challenges for 5G mobile antenna design and measurement. The ultrawide band (UWB) wireless system is another example of

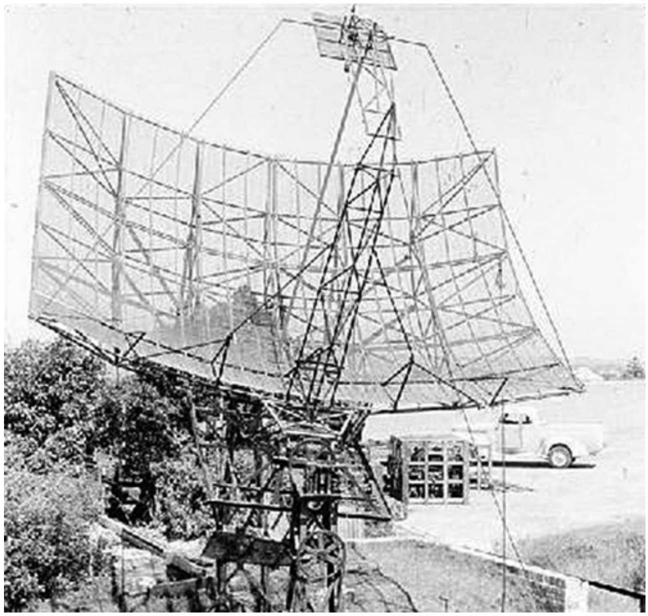

Figure 1.4 A WWII radar. Source: From ATNF, used with permission

recent broadband radio communication and positioning systems. The allocated frequency band is from 3.1 to 10.6 GHz. The beauty of UWB system is that the spectrum, which is normally very expensive, can be used free of charge but the power spectrum density is limited to −41.3 dBm/MHz. Thus, it is only suitable for short-distance applications (like Bluetooth but with a much larger bandwidth). The antenna design for these systems faces many challenging issues.

The role of antennas is becoming increasingly important. In some systems, the antenna is now no longer just a simple transmitting/receiving device, but a device which is integrated with other parts of the system to achieve better performance. For example, the MIMO antenna system has been introduced as an effective means to combat the multipath effects in the radio propagation channel and increase the channel capacity, where several co-ordinated antennas are required.

Things have been changing quickly in the wireless world. But one thing has never been changed since the very first antenna was made, that is, that the antenna is a practical engineering subject! It will remain as an engineering subject. Once an antenna is designed and made, it must be tested. How well it works is not just determined by the antenna itself, it also depends on the other parts of the system and the environment. The standalone antenna performance can be very different from that of an installed antenna. For example, when a mobile phone antenna is designed, we must take the case and other parts of the phone, even our hands, into account to ensure that it will work well in the real world. The antenna is an essential device of a radio system, but not an isolated device! This makes it an interesting and challenging subject.

1.2 Radio Systems and Antennas

A radio system is generally considered as an electronic system that employs radio waves, a type of EM wave up to GHz frequencies. An *antenna*, as an essential part of a radio system, is defined as a device that can radiate and receive EM energy in an efficient and desired manner. It is normally made of metal, but other materials may also be used. For example, ceramic materials have been employed to make dielectric resonator antennas (DRAs). There are many things in our lives, such as a power leads that can radiate and receive EM energy but cannot be viewed as antennas because the EM energy is not transmitted or received in an efficient and desired manner or because they are not a part of a radio system, thus they cannot be called antennas.

Since radio systems possess some unique and attractive advantages over wired systems, numerous radio systems have been developed. TV, radar, and mobile radio communication systems are just some examples. The advantages include at least:

- *Mobility*: it is essential for mobile communications;
- *Good coverage*: the radiation from an antenna can cover a very large area that is good for TV and radio broadcasting and mobile communications;
- *Low pathloss*: this is distance (and frequency) dependent. Since the loss of a transmission line is an exponential function of the distance (the loss in dB = distance × per unit loss in dB) and the loss of a radio wave is proportional to the distance square (the loss in dB = 20 $\log_{10}$ (distance)), thus the pathloss of radio waves can be much smaller than that of a cable link. For example, assume that the loss is 10 dB for both a transmission line and a radio wave over 100 m, if the distance is increased 10 times to 1000 m, the loss for the transmission line becomes $10 \times 10 = 100$ dB but the loss for the radio link is just $10 + 20 = 30$ dB, which is much smaller than 100 dB! Therefore, the radio communication system is extremely attractive for long-distance communication. It should be pointed out that optic fibers are also employed for long-distance communications since they are of very low loss and UWB – but it is for point-to-point communications and fibers/cables normally need to be buried in subsurface, which could be costly in practice.

Figure 1.5 illustrates a typical radio communication system. The source information is normally modulated and amplified in the transmitter and then passed on to the transmit antenna via a transmission line, which has a typical characteristic impedance (which will be explained in the Chapter 2) of 50 Ω. The antenna radiates the information in the form of an EM wave in an efficient and desired manner to the destination, where the information is picked up by the receiver antenna and passed on to the receiver via another transmission line. The signal is demodulated, and the original message is then recovered at the receiver.

Thus, the antenna is actually a transformer that transfers electrical signals (voltages and currents from a transmission line) into EM waves (electric and magnetic fields) or vice versa. For example, a satellite dish antenna receives the radio wave from a satellite and transfers it into electrical signals which are output to a cable to be further processed. Our eyes may be viewed as another example of antennas. In this case, the wave is not a radio wave but an optical wave, another form of EM wave that has much higher frequencies.

Now it is clear that the antenna is actually a transformer of voltage/current to electric/magnetic field; it can also be considered as a bridge to link the radio wave and transmission line. An *antenna system* is defined as the combination of the antenna and its feed line. As an antenna is

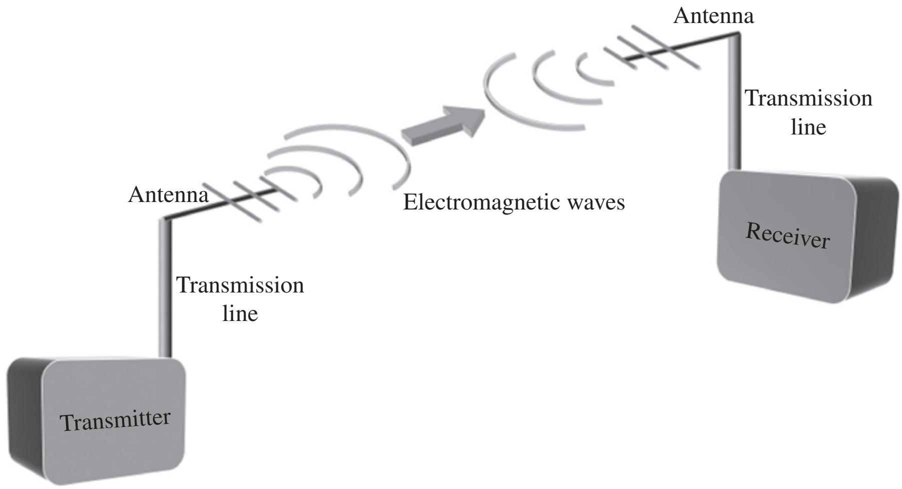

Figure 1.5 A typical radio system

usually connected to a transmission line, how to make this connection is a subject of interest since the signal from the feed line should be radiated into the space in an efficient and desired way. Transmission lines and radio waves are in fact two different subjects in engineering. To understand the antenna theory, one has to understand transmission lines and radio waves, which will be discussed in detail in Chapters 2 and 3, respectively. Thus, there is no need for the reader to study these subjects using other books.

In some applications where space is very limited (such as hand portables and aircrafts), it is desirable to integrate the antenna and its feed line. In other applications (such as the reception of TV broadcasting), the antenna is far away from the receiver and a long transmission line has to be used.

Unlike other devices in a radio system (such as filters and amplifiers), the antenna is a very special device; it deals with electrical signals (voltages and currents) as well as EM waves (electric fields and magnetic fields) that have made antenna design an interesting and difficult subject. For different applications, the requirements on the antenna could be very different, even for the same frequency band.

In conclusion, the subject of antennas is about how to design a suitable device that will be well matched with its feed line and radiate/receive the radio wave in an efficient and desired manner.

1.3 Necessary Mathematics

To thoroughly understand antenna theory requires a considerable amount of mathematics. However, the intention of this book is to provide the reader with a solid foundation of antenna theory and apply the theory to practical antenna design. Here, we are just going to introduce and review the essential and important mathematics required for this book. More in-depth study materials can be obtained from other references [1, 2].

1.3.1 Complex Numbers

In mathematics, a complex number, Z, consists of real and imaginary parts, that is

$$Z = R + jX \tag{1.1}$$

where R is called the real part of the complex number Z, i.e. Re(Z), and X is defined as the imaginary part of Z, i.e. Im(Z). Both R and X are real numbers, and j (not the traditional notion i in mathematics to avoid confusion with a changing current in electrical engineering) is the imaginary unit and defined by

$$j = \sqrt{-1} \tag{1.2}$$

Thus,

$$j^2 = -1 \tag{1.3}$$

Geometrically, a complex number can be presented in a two-dimensional plane where the imaginary part is found on the vertical axis while the real part is presented by the horizontal axis as shown in Figure 1.6.

In this model, multiplication by −1 corresponds to a rotation of 180 degrees about the origin. Multiplication by j corresponds to a 90-degree rotation anti-clockwise, and the equation $j^2 = -1$ is interpreted as saying that if we apply two 90-degree rotations about the origin, the net result is a single 180-degree rotation. Note that a 90-degree rotation clockwise also satisfies this interpretation.

Another representation of a complex number Z is to use the amplitude and phase form:

$$Z = Ae^{j\varphi} \tag{1.4}$$

where A is the amplitude and φ is the phase of the complex number Z, which are also shown in Figure 1.6. The two different representations are linked by the following equations:

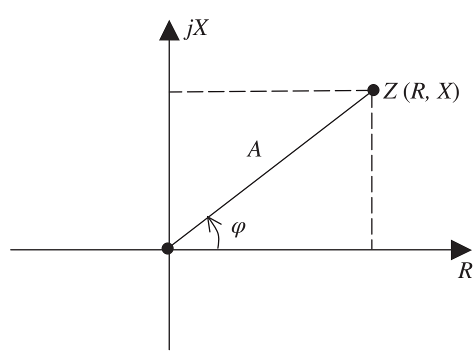

Figure 1.6 Complex plane

$$Z = R + jX = Ae^{j\varphi};$$
$$A = \sqrt{R^2 + X^2}, \qquad \varphi = \tan^{-1}(X/R) \tag{1.5}$$
$$R = A\cos\varphi, \qquad X = A\sin(\varphi)$$

1.3.2 Vectors and Vector Operation

A scalar is a one-dimensional quantity that has magnitude only, whereas a complex number is a two-dimensional quantity. A vector can be viewed as a three-dimensional (3D) quantity, and a special one – it has both a magnitude and a direction. For example, force and velocity are vectors. A position in space is a 3D quantity, but it does not have a direction, thus it is not a vector. Figure 1.7 is an illustration of vector $\boldsymbol{A}$ in Cartesian coordinates. It has three orthogonal components (A_x, A_y, A_z) along the x, y, and z directions, respectively. To distinguish vectors from scalars, the letter representing the vector is printed in bold, as $\boldsymbol{A}$ or $\boldsymbol{a}$, and a unit vector is printed in bold with a hat over the letter as $\hat{\boldsymbol{x}}$ or $\hat{\boldsymbol{n}}$.

The magnitude of vector $\boldsymbol{A}$ is given by

$$|\boldsymbol{A}| = A = \sqrt{A_x^2 + A_y^2 + A_z^2} \tag{1.6}$$

Now let us consider two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\boldsymbol{A} = A_x\hat{\boldsymbol{x}} + A_y\hat{\boldsymbol{y}} + A_z\hat{\boldsymbol{z}}$$
$$\boldsymbol{B} = B_x\hat{\boldsymbol{x}} + B_y\hat{\boldsymbol{y}} + B_z\hat{\boldsymbol{z}}$$

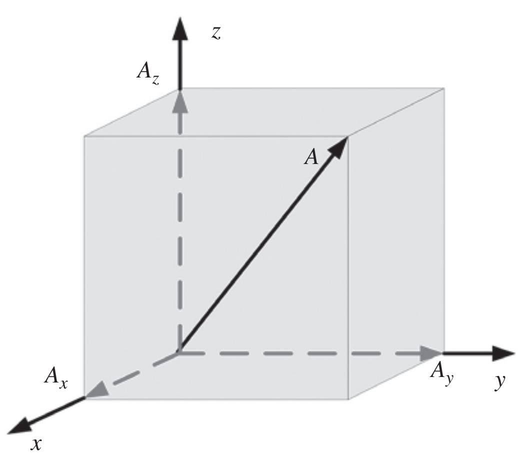

Figure 1.7 Vector A in Cartesian coordinates

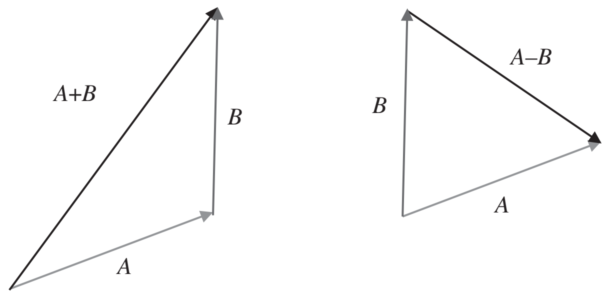

Figure 1.8 Vector addition and subtraction

The addition and subtraction of vectors can be expressed as

$$\begin{aligned}\boldsymbol{A}+\boldsymbol{B}&=(A_x+B_x)\hat{\boldsymbol{x}}+\left(A_y+B_y\right)\hat{\boldsymbol{y}}+(A_z+B_z)\hat{\boldsymbol{z}}\\ \boldsymbol{A}-\boldsymbol{B}&=(A_x-B_x)\hat{\boldsymbol{x}}+\left(A_y-B_y\right)\hat{\boldsymbol{y}}+(A_z-B_z)\hat{\boldsymbol{z}}\end{aligned}\tag{1.7}$$

Obviously, the addition obeys the *commutative law*, that is $\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{B}+\boldsymbol{A}$.

Figure 1.8 shows what the addition and subtraction mean geometrically. A vector may be multiplied or divided by a scalar. The magnitude changes but its direction remains the same. However, the multiplication of two vectors is complicated. There are two types of multiplication: dot product and cross product.

The *dot product* of two vectors is defined as

$$\boldsymbol{A}\bullet\boldsymbol{B}=|\boldsymbol{A}||\boldsymbol{B}|\cos\theta=A_xB_x+A_yB_y+A_zB_z\tag{1.8}$$

where θ is the angle between vector $\boldsymbol{A}$ and vector $\boldsymbol{B}$ and cos θ is also called the direction cosine. The dot • between $\boldsymbol{A}$ and $\boldsymbol{B}$ indicates the dot product that results in a scalar, thus it is also called a *scalar product*. If the angle θ is zero, $\boldsymbol{A}$ and $\boldsymbol{B}$ are in parallel – the dot product maximized, whereas for an angle of 90 degrees, i.e. when $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal, the dot product is zero.

It is worth noting that the dot product obeys the *commutative law*, that is, $\boldsymbol{A}\bullet\boldsymbol{B}=\boldsymbol{B}\bullet\boldsymbol{A}$.

The *cross product* of two vectors is defined as

$$\begin{aligned}\boldsymbol{A}\times\boldsymbol{B}&=\hat{\boldsymbol{n}}|\boldsymbol{A}||\boldsymbol{B}|\sin\theta=\boldsymbol{C}\\ &=\hat{\boldsymbol{x}}\left(A_yB_z-A_zB_y\right)+\hat{\boldsymbol{y}}\left(A_zB_x-A_xB_z\right)+\hat{\boldsymbol{z}}\left(A_xB_y-A_yB_x\right)\end{aligned}\tag{1.9}$$

where $\hat{\boldsymbol{n}}$ is a unit vector normal to the plane containing $\boldsymbol{A}$ and $\boldsymbol{B}$. The cross × between $\boldsymbol{A}$ and $\boldsymbol{B}$ indicates the cross product that results in a vector $\boldsymbol{C}$, thus, it is also called a *vector product*. The vector $\boldsymbol{C}$ is orthogonal to both $\boldsymbol{A}$ and $\boldsymbol{B}$, and the direction of $\boldsymbol{C}$ follows a so-called right-hand rule as shown in Figure 1.9. If the angle θ is zero or 180 degrees, that is, $\boldsymbol{A}$ and $\boldsymbol{B}$ are in parallel, the cross product is zero, whereas for an angle of 90 degrees, i.e. $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal, the cross

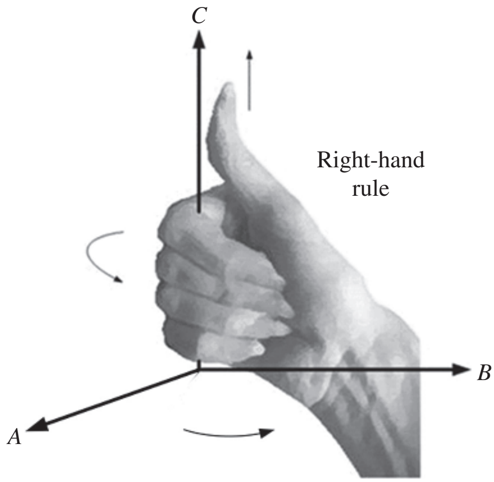

Figure 1.9 The cross product of vectors A and B

product of these two vectors reaches a maximum. Unlike the dot product, the cross product does not obey the commutative law.

The cross product may be expressed in determinant form as follows, which is the same as Equation (1.9) but it may be easier for some people to memorize:

$$\boldsymbol{A} \times \boldsymbol{B} = \begin{vmatrix} \hat{\boldsymbol{x}} & \hat{\boldsymbol{y}} & \hat{\boldsymbol{z}} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix} \tag{1.10}$$

Another important thing about vectors is that any vector can be decomposed into three orthogonal components (such as x, y, and z components) in 3D or two orthogonal components in a 2D plane.

Example 1.1 Vector operation

Vectors $\boldsymbol{A} = 10\hat{\boldsymbol{x}} + 5\hat{\boldsymbol{y}} + 1\hat{\boldsymbol{z}}$ and $\boldsymbol{B} = 2\hat{\boldsymbol{y}}$. Find:

$$\boldsymbol{A} + \boldsymbol{B}; \quad \boldsymbol{A} - \boldsymbol{B}; \quad \boldsymbol{A} \bullet \boldsymbol{B}; \quad \text{and } \boldsymbol{A} \times \boldsymbol{B}$$

Solution

$$\begin{aligned} \boldsymbol{A} + \boldsymbol{B} &= 10\hat{\boldsymbol{x}} + (5 + 2)\hat{\boldsymbol{y}} + 1\hat{\boldsymbol{z}} = 10\hat{\boldsymbol{x}} + 7\hat{\boldsymbol{y}} + 1\hat{\boldsymbol{z}}; \\ \boldsymbol{A} - \boldsymbol{B} &= 10\hat{\boldsymbol{x}} + (5 - 2)\hat{\boldsymbol{y}} + 1\hat{\boldsymbol{z}} = 10\hat{\boldsymbol{x}} + 3\hat{\boldsymbol{y}} + 1\hat{\boldsymbol{z}}; \\ \boldsymbol{A} \bullet \boldsymbol{B} &= 0 + (5 \times 2) + 0 = 10; \\ \boldsymbol{A} \times \boldsymbol{B} &= 10 \times 2\hat{\boldsymbol{z}} + 1 \times 2\hat{\boldsymbol{x}} = 20\hat{\boldsymbol{z}} + 2\hat{\boldsymbol{x}} \end{aligned}$$

1.3.3 Coordinates

In addition to the well-known Cartesian coordinates, the spherical coordinates (r, θ, φ), as shown in Figure 1.10, will also be used frequently throughout this book. These two coordinate systems have the following relations:

$$\begin{aligned} x &= r\sin\theta\cos\varphi \\ y &= r\sin\theta\sin\varphi \\ z &= r\cos\theta \end{aligned} \tag{1.11}$$

and

$$\begin{aligned} r &= \sqrt{x^2+y^2+z^2} \\ \theta &= \cos^{-1}\frac{z}{\sqrt{x^2+y^2+z^2}}; \quad 0 \le \theta \le \pi \\ \varphi &= \tan^{-1}\frac{y}{x}; \quad 0 \le \varphi \le 2\pi \end{aligned} \tag{1.12}$$

The dot products of unit vectors in these two coordinator systems are

$$\begin{aligned} &\hat{\boldsymbol{x}}\bullet\hat{\boldsymbol{r}} = \sin\theta\cos\varphi; \quad \hat{\boldsymbol{y}}\bullet\hat{\boldsymbol{r}} = \sin\theta\sin\varphi; \quad \hat{\boldsymbol{z}}\bullet\hat{\boldsymbol{r}} = \cos\theta \\ &\hat{\boldsymbol{x}}\bullet\hat{\boldsymbol{\theta}} = \cos\theta\cos\varphi; \quad \hat{\boldsymbol{y}}\bullet\hat{\boldsymbol{\theta}} = \cos\theta\sin\varphi; \quad \hat{\boldsymbol{z}}\bullet\hat{\boldsymbol{\theta}} = -\sin\theta \\ &\hat{\boldsymbol{x}}\bullet\hat{\varphi} = -\sin\varphi; \quad \hat{\boldsymbol{y}}\bullet\hat{\varphi} = \cos\varphi; \quad \hat{\boldsymbol{z}}\bullet\hat{\varphi} = 0 \end{aligned} \tag{1.13}$$

Thus, we can express a quantity in one coordinate system using the known parameters in the other coordinate system. For example, if A_r, A_θ, A_φ are known, we can find

$$A_x = \boldsymbol{A}\bullet\hat{\boldsymbol{x}} = A_r\sin\theta\cos\varphi + A_\theta\cos\theta\cos\varphi - A_\varphi\sin\varphi$$

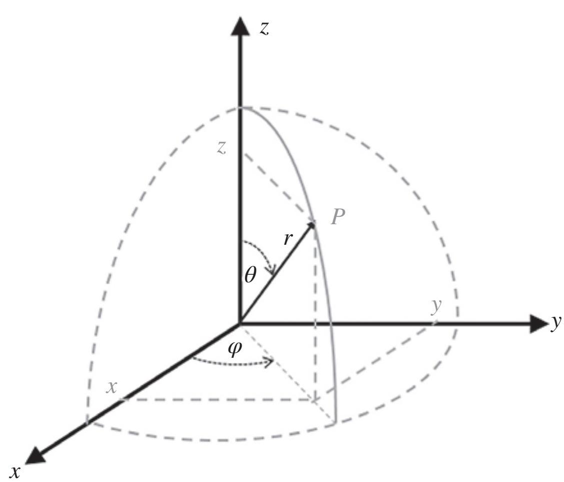

Figure 1.10 Cartesian and spherical coordinates

1.4 Basics of EMs

Now let us use basic mathematics to deal with antennas, or precisely, EM problems in this section.

EM waves cover the whole spectrum; radio waves and optical waves are just two examples of EM waves. We can see light but cannot see radio waves. The whole spectrum is divided into many frequency bands. Some EM bands and their applications are listed in Table 1.1. There are

Table 1.1 EM frequency bands and applications

Frequency	Band	Wavelength	Applications
3–30 kHz	VLF	100–10 km	Navigation, sonar, fax
30–300 kHz	LF	10–1 km	Navigation
0.3–3 MHz	MF	1–0.1 km	AM broadcasting
3–30 MHz	HF	100–10 m	Tel, Fax, CB, ship communications
30–300 MHz	VHF	10–1 m	TV, FM broadcasting
0.3–3 GHz	UHF	1–0.1 m	TV, mobile, radar
3–30 GHz	SHF	100–10 mm	Radar, satellite, mobile, microwave links
30–300 GHz	EHF	10–1 mm	Radar, wireless communications
0.3–3 THz	THz	1–0.1 mm	THz imaging
3–430 THz	Infrared	0.1 mm–700 nm	Heating, communications, camera
430–770 THz	Light	700–400 nm	Lighting, camera
Radar frequency bands according to IEEE standard			
1–2 GHz	L	0.3–0.15 m	Long wave, mobile radio
2–4 GHz	S	0.15–0.075 m	Short wave, mobile radio
4–8 GHz	C	7.5–3.75 cm	Compromise between S and X, radar
8–12 GHz	X	3.75–2.5 cm	Radar, satellite
12–18 GHz	Ku	2.5–1.7 cm	Satellite and radar
18–27 GHz	K	1.7–1.1 cm	Satellite and radar
27–40 GHz	Ka	11–7.5 mm	Communications and radar
40–75 GHz	V	7.5–4.0 mm	Communications and radar
75–110 GHz	W	4.0–2.7 mm	Communications and radar
Frequencies of some popular wireless systems			
535–1605 kHz	AM radio broadcast band		
3–30 MHz	Short-wave radio broadcast band		
13.56 MHz	NFC		
88–108 MHz	FM radio broadcast band		
175–240 MHz	DAB radio broadcast band		
470–890 MHz	UHF TV (14-83)		
698–960 MHz	Cellular mobile radio (2/4G)		
1710–2690 MHz	Cellular mobile radio (2/3/4G)		
3.3–3.8 GHz	Cellular mobile radio (5G)		
1.227 GHz	GPS L2 band		
1.575 GHz	GPS L1 band		
2.45 GHz	Microwave, Bluetooth, Wi-Fi		
3.1–10.6 GHz	UWB band		
5.180–5.825	Wi-Fi bands		

other letter band designations from organizations such as NATO. Here, we have used the IEEE standard.

Although the whole spectrum is infinite, the useful spectrum is limited and some frequency bands, such as the UHF, are already very congested. Normally significant license fees have to be paid to use the spectrum, although there are some license-free bands: the most well-known ones are the industrial, science, and medical (ISM) bands. The 433 MHz and 2.45 GHz bands are just two examples. Cable operators do not need to pay the spectrum license fees, but they have to pay other fees for things such as digging out the roads to bury the cables.

The *wave velocity* v is linked to the frequency f and wavelength λ by this simple equation:

$$v = f\lambda \tag{1.14}$$

It is well known that the speed of light (an EM wave) is about 3×10^8 m/s in free space. The higher the frequency, the shorter the wavelength. An illustration of how the frequency is linked to the wavelength is given in Figure 1.11, where both the frequency and wavelength are plotted on a logarithmic scale. The advantage, by doing so, is that we can see clearly how the function is changed even over a very large scale.

Radio waves, lights, and X-ray ($f = 10^{16}$ to 10^{19} Hz) are EM waves at different frequencies although they seem to be very different. One thing that all the forms of EM waves have in common is that they can travel through empty space (vacuum). This is not true for other kinds of waves; sound waves, for example, need some kind of material, such as air or water, in which to move. EM energy is carried by photons, the energy of a photon (also called quantum energy) is hf, where h is Planck's constant $= 6.63 \times 10^{-34}$ Js, and f is frequency in Hz. The higher the frequency, the more the energy of a photon. X-Ray has been used for imaging just because

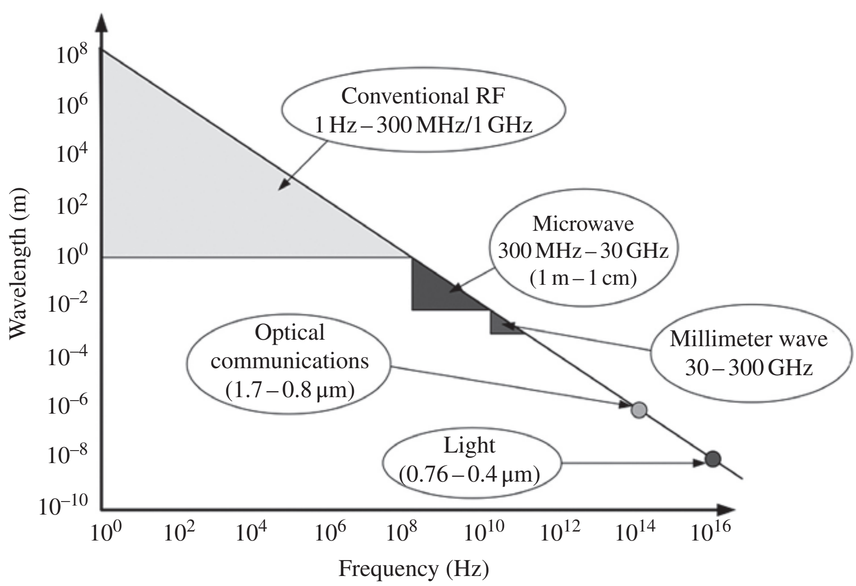

Figure 1.11 Frequency vs wavelength

of its high frequency: it carries very high energy and can penetrate through most objects. Also due to this high energy, X-ray can kill our cells and cause ionizing radiation that is not safe for our health. However, lights and radio waves operate at lower frequencies and do not have such a problem.

Logarithmic scales are widely used in RF (radio frequency) engineering and antennas community since the signals we are dealing with change significantly (over 1000 times in many cases) in terms of the magnitude. The signal power is normally expressed in *dB* (decibel), which is defined as

$$P\,(\text{dBW}) = 10\log_{10}\frac{P\,(\text{W})}{1\,\text{W}}; \quad P\,(\text{dBm}) = 10\log_{10}\frac{P\,(\text{W})}{1\,\text{mW}} \tag{1.15}$$

Thus, 100 W is 20 dBW, or just expressed as 20 dB in most cases; 1 W is 0 dB or 30 dBm; and 0.5 W is −3 dB or 27 dBm. Based on this definition, we can also express other parameters in dB. For example, since the power is linked to voltage V by $P = V^2/R$ (so $P \propto V^2$), the voltage can be converted to dBV by

$$V\,(\text{dBV}) = 20\log_{10}\left(\frac{V\,(\text{V})}{1\,\text{V}}\right) \tag{1.16}$$

Thus, 300 kVolts is 70 dBV, and 0.5 V is −6 dBV (not −3 dBV) or 54 dBmV.

1.4.1 Electric Field

The *electric field* (in V/m) is defined as the force (in Newtons) per unit charge (in Coulombs). From this definition and Coulomb's law, the electric field $\boldsymbol{E}$ created by a single point charge Q at a distance r is

$$\boldsymbol{E} = \frac{\boldsymbol{F}}{Q} = \frac{Q}{4\pi\varepsilon r^2}\hat{\boldsymbol{r}}\ (\text{V/m}) \tag{1.17}$$

where

$\boldsymbol{F}$ is the electric force given by Coulomb's law $\left(\boldsymbol{F} = \dfrac{Q_1 Q_2}{4\pi\varepsilon r^2}\hat{\boldsymbol{r}}\right)$;

$\hat{\boldsymbol{r}}$ is a unit vector along $\boldsymbol{r}$ direction which is also the direction of the electric field $\boldsymbol{E}$.

ε is the electric *permittivity* of the material. Its SI unit is Farads/m. In free space, it is a constant:

$$\varepsilon_0 = 8.85419 \times 10^{-12}\ \text{F/m} \tag{1.18}$$

The product of permittivity and the electric field is called the *electric flex density,* $\boldsymbol{D}$, which is a measure of how much electric flux passes through a unit area, i.e.

$$\boldsymbol{D} = \varepsilon\boldsymbol{E} = \varepsilon_r\varepsilon_0\boldsymbol{E}\ \left(\text{C/m}^2\right) \tag{1.19}$$

where $\varepsilon_r = \varepsilon/\varepsilon_0$ is called the *relative permittivity* (also called *dielectric constant*, but it is normally a function of frequency, not really a constant, thus relative permittivity is preferred in this

Table 1.2 Relative permittivity of some common materials at 100 MHz

Material	Relative permittivity	Material	Relative permittivity
ABS (plastic)	2.4–3.8	Polypropylene	2.2
Air	1	Polyvinylchloride (PVC)	3
Alumina	9.8	Porcelain	5.1–5.9
Aluminum silicate	5.3–5.5	PTFE-teflon	2.1
Balsa wood	1.37 @ 1 MHz	PTFE-ceramic	10.2
	1.22 @ 3 GHz	PTFE-glass	2.1–2.55
Concrete	~8	RT/Duroid 5870	2.33
Copper	1	RT/Duroid 6006	6.15 @ 3 GHz
Diamond	5.5–10	Rubber	3.0–4.0
Epoxy (FR4)	4.4	Sapphire	9.4
Epoxy glass PCB	5.2	Sea water	80
Ethyl alcohol (absolute)	24.5 @ 1 MHz	Silicon	11.7–12.9
	6.5 @ 3 GHz	Soil	~10
FR-4(G-10)		Soil (dry sandy)	2.59 @ 1 MHz
– low resin	4.9	Water (32 °F)	88.0
– high resin	4.2	(68 °F)	80.4
GaAs	13.0	(212 °F)	55.3
Glass	~4	Wood	~2
Gold	1		
Ice (pure distilled water)	4.15 @ 1 MHz		
	3.2 @ 3 GHz		

book). The relative permittivities of some common materials are listed in Table 1.2. Note that they are functions of frequency and temperature. Normally, the higher the frequency, the smaller the permittivity in the radio frequency band. It should also be pointed out that almost all conductors have a relative permittivity of one.

The electric flux density is also called the *electric displacement*, hence, the symbol $\boldsymbol{D}$. It is also a vector. In an isotropic material (properties independent of direction) $\boldsymbol{D}$ and $\boldsymbol{E}$ are in the same direction and ε is a scalar quantity. In an anisotropic material, $\boldsymbol{D}$ and $\boldsymbol{E}$ may be in different directions if ε is a tensor.

If the permittivity is a complex number, it means that the material has some loss. The *complex permittivity* can be written as

$$\varepsilon = \varepsilon' - j\varepsilon'' \tag{1.20}$$

The ratio of the imaginary part to the real part is called the *loss tangent*, that is

$$\tan\delta = \frac{\varepsilon''}{\varepsilon'} \tag{1.21}$$

It has no unit and is also a function of frequency and temperature.

Table 1.3 Conductivities of some common materials at room temperature

Material	Conductivity (S/m)	Material	Conductivity (S/m)
Silver	6.3×10^7	Graphite	$\approx 10^5$
Copper	5.8×10^7	Carbon	$\approx 10^4$
Gold	4.1×10^7	Silicon	$\approx 10^3$
Aluminum	3.5×10^7	Ferrite	$\approx 10^2$
Tungsten	1.8×10^7	Sea water	≈ 5
Zinc	1.7×10^7	Germanium	≈ 2
Brass	1×10^7	Wet soil	≈ 1
Phosphor bronze	1×10^7	Animal blood	0.7
Tin	9×10^6	Animal body	0.3
Lead	5×10^6	Fresh water	$\approx 10^{-2}$
Silicon steel	2×10^6	Dry soil	$\approx 10^{-3}$
Stainless steel	1×10^6	Distilled water	$\approx 10^{-4}$
Mercury	1×10^6	Glass	$\approx 10^{-12}$
Cast iron	$\approx 10^6$	Air	0

The electric field $\boldsymbol{E}$ is related to the current density $\boldsymbol{J}$ (in A/m^2), another important parameter, by Ohm's law. The relationship between them at a point can be expressed as

$$\boldsymbol{J} = \sigma \boldsymbol{E} \tag{1.22}$$

where σ is the *conductivity*, which is the reciprocal of *resistivity*. It is a measure of a material's ability to conduct an electrical current and is expressed in Siemens per meter (S/m). Table 1.3 lists conductivities of some common materials linked to antenna engineering. The conductivity is also a function of temperature and frequency.

1.4.2 Magnetic Field

Whilst charges can generate an electric field, currents can generate a magnetic field. The *magnetic field*, $\boldsymbol{H}$ (in A/m), is the vector field that forms closed loops around electric currents or magnets. The magnetic field from a current vector $\boldsymbol{I}$ is given by the Biot–Savart law as

$$\boldsymbol{H} = \frac{\boldsymbol{I} \times \hat{\boldsymbol{r}}}{4\pi r^2} \text{ (A/m)} \tag{1.23}$$

where

$\hat{\boldsymbol{r}}$ is the unit displacement vector from the current element to the field point and
r is the distance from the current element to the field point.
$\boldsymbol{I}$, $\hat{\boldsymbol{r}}$ and $\boldsymbol{H}$ follow the right-hand rule, that is $\boldsymbol{H}$ is orthogonal to both $\boldsymbol{I}$ and $\hat{\boldsymbol{r}}$, as illustrated by Figure 1.12.

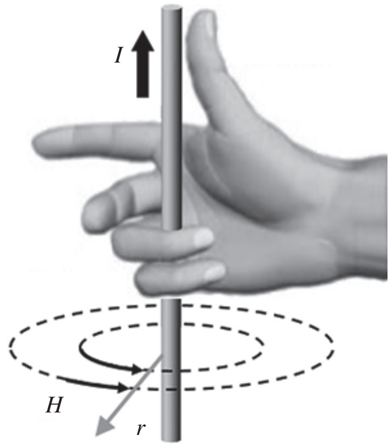

Figure 1.12 Magnetic field generated by current ***I***

Like the electric field, the magnetic field exerts a force on electric charge. But unlike an electric field, it employs force only on a moving charge, and the direction of the force is orthogonal to both the magnetic field and the charge's velocity:

$$\boldsymbol{F} = Q\boldsymbol{v} \times \mu\boldsymbol{H} \tag{1.24}$$

where

$\boldsymbol{F}$ is the force vector produced, measured in Newtons;
Q is the electric charge that the magnetic field is acting on, measured in Coulombs (C);
$\boldsymbol{v}$ is the velocity vector of the electric charge Q, measured in meters per second (m/s);
μ is the magnetic permeability of the material. Its unit is Henries per meter (H/m). The permeability of free space is

$$\mu_0 = 4\pi \times 10^{-7}\ \mathrm{H/m} \tag{1.25}$$

In Equation (1.24), $Q\boldsymbol{v}$ can actually be viewed as the current vector $\boldsymbol{I}$ and the product of $\mu\boldsymbol{H}$ is called the *magnetic flux density* $\boldsymbol{B}$ (in Tesla), the counterpart of the electric flux density. Thus,

$$\boldsymbol{B} = \mu\boldsymbol{H} \tag{1.26}$$

Again, in an isotropic material (properties independent of direction), $\boldsymbol{B}$ and $\boldsymbol{H}$ are in the same direction and μ is a scalar quantity. In an anisotropic material, $\boldsymbol{B}$ and $\boldsymbol{E}$ may be in different directions and μ is a tensor.

Like the relative permittivity, the relative permeability is given as

$$\mu_r = \mu/\mu_0 \tag{1.27}$$

Table 1.4 Relative permeabilities of some common materials

Material	Relative permeability	Material	Relative permeability
Superalloy	$\approx 1 \times 10^6$	Aluminum	≈ 1
Purified iron	$\approx 2 \times 10^5$	Air	1
Silicon iron	$\approx 7 \times 10^3$	Water	≈ 1
Iron	$\approx 5 \times 10^3$	Copper	≈ 1
Mild steel	$\approx 2 \times 10^3$	Lead	≈ 1
Nickel	600	Silver	≈ 1

The relative permeabilities of some materials are given in Table 1.4. Permeability is not sensitive to frequency and temperature. Most materials, including conductors, have a relative permeability very close to one.

Combining Equations (1.17) and (1.24) yields

$$\boldsymbol{F} = Q(\boldsymbol{E} + \boldsymbol{v} \times \mu \boldsymbol{H}) \tag{1.28}$$

This is called the *Lorentz force*. The particle will experience a force due to the electric field of $Q\boldsymbol{E}$ and the magnetic field $Q\boldsymbol{v} \times \boldsymbol{B}$.

1.4.3 Maxwell's Equations

Maxwell's equations are a set of equations first presented as a distinct group in the latter half of the nineteenth century by James Clerk Maxwell (1831–1879) pictured in Figure 1.13. Mathematically, they can be expressed in the following differential form:

$$\begin{aligned}
\nabla \times \boldsymbol{E} &= -\frac{d\boldsymbol{B}}{dt} \\
\nabla \times \boldsymbol{H} &= \boldsymbol{J} + \frac{d\boldsymbol{D}}{dt} \\
\nabla \bullet \boldsymbol{D} &= \rho \\
\nabla \bullet \boldsymbol{B} &= 0
\end{aligned} \tag{1.29}$$

where

ρ is the charge density.

$\nabla = \frac{\partial}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial}{\partial z}\hat{\boldsymbol{z}}$ is a vector operator;

$\nabla\times$ is the *curl operator* or called *rot* in some countries instead of *curl*.

$\nabla\bullet$ is the *divergence operator*.

Here we have both the vector cross product and the dot product.

Maxwell's equations describe the interrelationship between electric fields, magnetic fields, electric charge, and electric current. Although Maxwell himself was not the originator of the individual equations, he derived them again independently in conjunction with his molecular vortex model of Faraday's lines of force, and he was the person who first grouped these

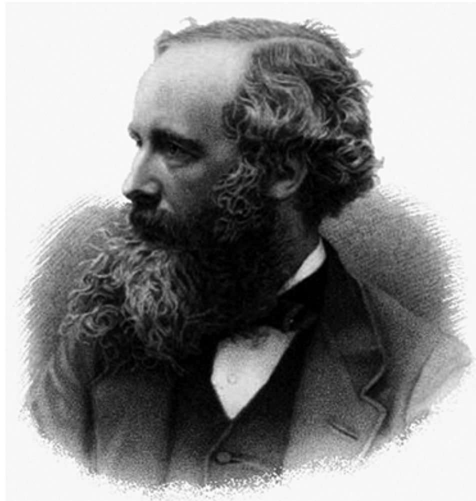

Figure 1.13 James Clerk Maxwell

equations all together into a coherent set. Most importantly, he introduced an extra term to Ampere's Circuital Law, the second equation of (1.19). This extra term is the time derivative of the electric field and is known as Maxwell's displacement current. Maxwell's modified version of Ampere's Circuital Law enables the set of equations to be combined together to derive the EM wave equation, which will be further discussed in Chapter 3.

Now let us have a closer look at these mathematical equations to see what they really mean in terms of the physical explanations.

1.4.3.1 Faraday's Law of Induction

$$\nabla \times \boldsymbol{E} = -\frac{d\boldsymbol{B}}{dt} \tag{1.30}$$

This equation simply means that the induced *electromotive force* (EMF with a unit in V, it is the left-hand size of the equation expressed in the integral form as shown in Equation (1.37)) is proportional to the rate of change of the magnetic flux. In layman's terms, moving a conductor (such as a metal wire) through a magnetic field produces a voltage. The resulting voltage is directly proportional to the speed of movement. It is apparent from this equation that a time-varying magnetic field $\left(\mu \frac{d\boldsymbol{H}}{dt} \neq 0\right)$ will generate an electric field, i.e. $\boldsymbol{E} \neq 0$. But if the magnetic field is not time varying, it will NOT generate an electric field!

1.4.3.2 Amperes' Circuital Law

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{d\boldsymbol{D}}{dt} \tag{1.31}$$

This equation was modified by Maxwell by introducing the displacement current $\frac{d\boldsymbol{D}}{dt}$. It means that a magnetic field appears during the charge or discharge of a capacitor. With this concept, and the Faraday's law, Maxwell was able to derive the wave equations, and by showing that the predicted wave velocity was the same as the measured velocity of light, Maxwell asserted that light waves are EM waves.

This equation shows that both the current ($\boldsymbol{J}$) and time-varying electric field $\left(\varepsilon \frac{d\boldsymbol{E}}{dt}\right)$ can generate a magnetic field, i.e. $\boldsymbol{H} \neq 0$.

1.4.3.3 Gauss' Law for Electric Field

$$\nabla \bullet \boldsymbol{D} = \rho \tag{1.32}$$

This is the electrostatic application of Gauss's generalized theorem, giving the equivalence relation between any flux, e.g. of liquids, electric or gravitational, flowing out of any closed surface and the result of inner sources and sinks, such as electric charges or masses enclosed within the closed surface. As a result, it is not possible for electric fields to form a closed loop. Since $\boldsymbol{D} = \varepsilon\boldsymbol{E}$, it is also clear that charges (ρ) can generate electric fields, i.e. $\boldsymbol{E} \neq 0$.

1.4.3.4 Gauss' Law for Magnetic Field

$$\nabla \bullet \boldsymbol{B} = 0 \tag{1.33}$$

This shows that the divergence of the magnetic field ($\nabla \bullet \boldsymbol{B}$) is always zero, which means that the magnetic field lines are closed loops, thus the integral of $\boldsymbol{B}$ over a closed surface is zero.

For a time-harmonic EM field (which means the field linked to the time by factor $e^{j\omega t}$ where ω is the angular frequency and t is the time), we can use the *constitutive relations*

$$\boldsymbol{D} = \varepsilon\boldsymbol{E}, \quad \boldsymbol{B} = \mu\boldsymbol{H}, \quad \boldsymbol{J} = \sigma\boldsymbol{E} \tag{1.34}$$

to write Maxwell's equations into the following forms

$$\begin{aligned} \nabla \times \boldsymbol{E} &= -j\omega\mu\boldsymbol{H} \\ \nabla \times \boldsymbol{H} &= \boldsymbol{J} + j\omega\varepsilon\boldsymbol{E} = j\omega\varepsilon\left(1 - j\frac{\sigma}{\omega\varepsilon}\right)\boldsymbol{E} \\ \nabla \bullet \boldsymbol{E} &= \rho/\varepsilon \\ \nabla \bullet \boldsymbol{H} &= 0 \end{aligned} \tag{1.35}$$

where $\boldsymbol{B}$ and $\boldsymbol{D}$ are replaced by the electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{H}$ to simplify the equations and they will not appear again unless necessary.

It should be pointed out that, in Equation (1.35), $\varepsilon\left(1-j\dfrac{\sigma}{\omega\varepsilon}\right)$ can be viewed as a *complex permittivity* defined by Equation (1.20). In this case, the loss tangent is

$$\tan\delta = \frac{\varepsilon''}{\varepsilon'} = \frac{\sigma}{\omega\varepsilon} \tag{1.36}$$

It is hard to predict how the loss tangent changes with the frequency since both the permittivity and conductivity are functions of frequency as well. More discussion will be given in Chapter 3.

1.4.4 Boundary Conditions

Maxwell's equations can also be written in the integral form as

$$\begin{aligned}
&\oint_C \boldsymbol{E}\bullet d\boldsymbol{l} = -\iint_S \frac{d\boldsymbol{B}}{dt}\bullet d\boldsymbol{s}\\
&\oint_C \boldsymbol{H}\bullet d\boldsymbol{l} = \iint_S \left(\boldsymbol{J} + \frac{d\boldsymbol{D}}{dt}\right)\bullet d\boldsymbol{s}\\
&\oiint_S \boldsymbol{D}\bullet d\boldsymbol{s} = \iiint_V \rho dv = Q\\
&\oiint_S \boldsymbol{B}\bullet d\boldsymbol{s} = 0
\end{aligned} \tag{1.37}$$

Consider the boundary between two materials shown in Figure 1.14. Using these equations, we can obtain a number of useful results. For example, if we apply the first equation of Maxwell's equations in integral form to the boundary between Medium 1 and Medium 2, it is not difficult to obtain [2]:

$$\hat{\boldsymbol{n}} \times \boldsymbol{E}_1 = \hat{\boldsymbol{n}} \times \boldsymbol{E}_2 \tag{1.38}$$

where $\hat{\boldsymbol{n}}$ is the surface unit vector from Medium 2 to Medium 1 as shown in Figure 1.14. This condition means that the tangential components of an electric field ($\hat{\boldsymbol{n}} \times \boldsymbol{E}$) are continuous across the boundary between any two media.

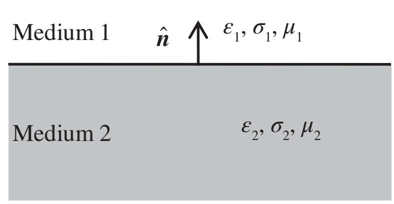

Figure 1.14 Boundary between Medium 1 and Medium 2

Similarly, we can apply other three Maxwell's equations to this boundary to obtain:

$$\begin{aligned} &\hat{\boldsymbol{n}} \times (\boldsymbol{H}_1 - \boldsymbol{H}_2) = \boldsymbol{J}_s \\ &\hat{\boldsymbol{n}} \bullet (\varepsilon_1 \boldsymbol{E}_1 - \varepsilon_2 \boldsymbol{E}_2) = \rho_s \\ &\hat{\boldsymbol{n}} \bullet (\mu_1 \boldsymbol{H}_1 - \mu_2 \boldsymbol{H}_2) = 0 \end{aligned} \tag{1.39}$$

where $\boldsymbol{J}_s$ is the surface current density and ρ_s is the surface charge density. These results can be interpreted as

- The change in tangential component of the magnetic field across a boundary is equal to the surface current density on the boundary;
- The change in the normal component of the electric flux density across a boundary is equal to the surface charge density on the boundary;
- The normal component of the magnetic flux density is continuous across the boundary between two media, while the normal component of the magnetic field is not continuous unless $\mu_1 = \mu_2$.

Applying these boundary conditions on a perfect conductor (which means no electric and magnetic field inside and the conductivity $\sigma = \infty$) in the air, we have

$$\hat{\boldsymbol{n}} \times \boldsymbol{E} = 0; \quad \hat{\boldsymbol{n}} \times \boldsymbol{H} = \boldsymbol{J}_s; \quad \hat{\boldsymbol{n}} \bullet \boldsymbol{E} = \rho_s/\varepsilon; \quad \hat{\boldsymbol{n}} \bullet \boldsymbol{H} = 0 \tag{1.40}$$

We can also use these results to illustrate, for example, the field distribution around a two-wire transmission line as shown in Figure 1.15, where the electric fields are plotted as the solid lines and the magnetic fields are shown in broken lines. As expected, the electric field is from positive charges to the negative charges, while the magnetic field forms loops around the current.

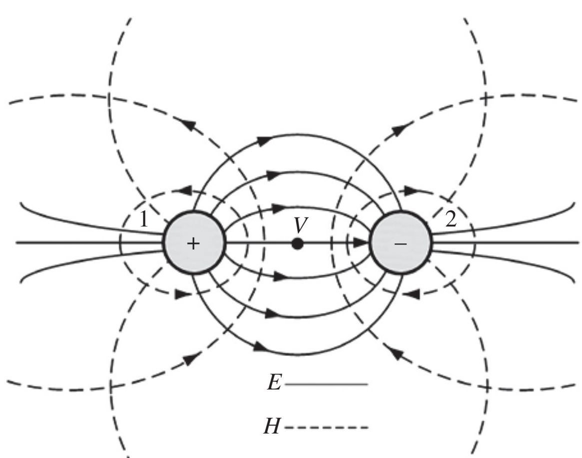

Figure 1.15 Electromagnetic field distribution around a two-wire transmission line

1.5 Summary

In this chapter, we have introduced the concept of antennas, briefly reviewed antenna history, and laid down the mathematical foundations for further study. The focus has been on the basics of EMs, which include electric and magnetic fields, EM properties of materials, Maxwell's equations, and boundary conditions. Maxwell's equations have revealed how electric fields, magnetic fields, and sources (currents and charges) are interlinked. They are the foundation of EMs and antennas.

References

1. R. E. Collin, Antennas and Radiowave Propagation, McGraw-Hill, Inc., 1985.
2. J. D. Kraus and D. A. Fleisch, Electromagnetics with Applications, 5th edition, McGraw-Hill, Inc., 1999.

Problems

Q1.1. What wireless communication experiment did H. Hertz conduct in 1887? Use a diagram to illustrate your answer.

Q1.2. Use an example to explain what a complex number means in our daily life.

Q1.3. Vector $\boldsymbol{A} = 10\hat{\boldsymbol{x}} + 5\hat{\boldsymbol{y}} + 1\hat{\boldsymbol{z}}$ and $\boldsymbol{B} = 5\hat{\boldsymbol{z}}$. Find
- a. the amplitude of vector ***A;***
- b. the angle between vectors ***A*** and ***B***;
- c. the dot product of these two vectors;
- d. a vector which is orthogonal to ***A*** and ***B***.

Q1.4. Vector $\boldsymbol{A} = 10\sin(10t + 10z)\hat{\boldsymbol{x}} + 5\hat{\boldsymbol{y}}$. Find
- a. $\nabla \bullet \mathbf{A}$;
- b. $\nabla \times \mathbf{A}$;
- c. $(\nabla \bullet \nabla)\mathbf{A}$;
- d. $\nabla \nabla \bullet \mathbf{A}$

Q1.5. Vector $\boldsymbol{E} = 10e^{j(10t-10z)}\hat{\boldsymbol{x}}$. Find
- a. The amplitude of ***E***;
- b. Plot the real part of ***E*** as a function of *t*;
- c. Plot the real part of ***E*** as a function of *z*;
- d. What this vector means.

Q1.6. Explain why mobile phone service providers have to pay license fees to us the spectrum. Who is responsible for the spectrum allocation in your country?

Q1.7. Cellular mobile communications have become part of our daily life. Explain the major differences between the 2nd, 3rd, and 4th generations of cellular mobile systems in terms of the frequency, data rate, and bandwidth. Further explain why their operational frequencies have increased.

Q1.8. Which frequency bands have been used for radar applications? Give an example.

Q1.9. Express 1 kW in dB, 10 kV in dBV, 0.5 dB in W, and 40 dBμV/m in V/m and μV/m.

Q1.10. Explain the concepts of the electric field and magnetic field, how are they linked to the electric and magnetic flux density functions?

Q1.11. What are the material properties of interest to our electromagnetic and antenna engineers?

Q1.12. What is the Lorentz force? Name an application of the Lorentz force in our daily life.

Q1.13. If a magnetic field on a conducting surface $z = 0$ is $\boldsymbol{H} = 10\cos(10t - 5z)\hat{\boldsymbol{x}}$, find the surface current density $\boldsymbol{J_s}$.

Q1.14. Use Maxwell's equations to explain the major differences between the static EM fields and time-varying EM fields.

Q1.15. Express the boundary conditions for the electric and magnetic fields on the surface of a perfect conductor.