Lectures on Electromagnetic Field Theory

Weng Cho Chew

Fall 2023,[1] Purdue University

[1]Updated March 18, 2024

DEDICATED TO THE MEMORY OF

ANDREW M. WEINER

Preface

This set of lecture notes is from my teaching of Electromagnetic Field Theory, at ECE, Purdue University, West Lafayette. It is intended for entry level graduate students. Because different universities have different undergraduate requirements in electromagnetic field theory, this is a course intended to "level the playing field". From this point onward, hopefully, all students will have the fundamental background in electromagnetic field theory needed to take advance level courses and do research at Purdue.

In developing this course, I have drawn heavily upon knowledge of our predecessors in this area. Many of the textbooks and papers used, I have listed them in the reference list. Being a practitioner in this field for over 40 years, I have seen electromagnetic theory impacting modern technology development unabated. Despite its age, the set of Maxwell's equations has enduring legacy and continued to be important, from statics to optics (EUV), from classical to quantum, and from sub-nanometer (subatomic) lengthscales to galactic lengthscales. The applications of electromagnetic technologies have also been tremendous and wide-ranging: from geophysical exploration, remote sensing, bio-sensing, electrical machinery, renewable and clean energy, biomedical engineering, optics and photonics, computer chip design, computer system, quantum computer designs, quantum communication and many more. Electromagnetic field theory is not everything, but it remains an indispensable component of modern technology developments.

The challenge in teaching this course is on how to teach over 150 years of knowledge in one semester: Of course this is mission impossible! To do this, we use the traditional wisdom of engineering education: Distill the knowledge, make it as simple as possible, and teach the fundamental big ideas in one short semester. Because of this, you may find the flow of the lectures erratic and rapid. Sometimes, I feel the need to touch on certain big ideas before moving on, resulting in the choppiness of the curriculum.

Also, in this course, I exploit mathematical "homomorphism" as much as possible to simplify the teaching. After years of practising in this area, I find that many complex and advanced concepts become simpler if mathematical homomorphism is exploited between the advanced concepts and simpler ones. An example of this is on waves in layered media. The problem is homomorphic to the transmission line problem: Hence, using transmission line theory, one can simplify the derivations of some complicated formulas.

A large part of modern applied electromagnetic technologies is based on heuristics. This is something difficult to teach, as it relies on physical insight and experience. Modern commercial software has reshaped this landscape: The field of math-physics modeling through numerical simulations, known as computational electromagnetic (CEM), has made rapid advances in recent years. Many cut-and-try laboratory experiments, based on heuristics, have been replaced by cut-and-try

computer experiments, which are a lot cheaper.

An exciting modern development is the role of electromagnetics and Maxwell's equations in quantum technologies. We will connect Maxwell's equations to quantum electromagnetics toward the end of this course. This is a challenge, as it has never been done before at an entry level course to my knowledge.

The first vacuum tube computer, ENIAC was built around 1945. After that, in the 1950s, a series of vacuum tube plus transistor computers were built including the ILLIAC series at U of Illinois. Those computers can fill a whole room. After some 70 years, with the compounding effect of nanotechnologies, we can now carry a pocket-size cell phone packed with billions of transistors. A change in *modus operandi* is that engineering designs are done increasingly more with software to reduce cost rather than cut-and-try experiments. Thus, an important field of computational electromagnetics (CEM) has emerged in recent years. Virtual prototyping of engineering designs can be done with software rather than hardware. In fact, 95 percent of a computer chip design is now done with software simulation greatly reducing the design cost. Unfortunately, we can only spend two lectures on CEM to convey some of the big ideas across to the students of electromagnetics. The devil is in the details in the implementations of these big ideas, which can be pursued in other courses.

Weng Cho Chew
March 18, 2024 Purdue University

Acknowledgements

I like to thank Dan Jiao for sharing her lecture notes in this course, as well as Andy Weiner for sharing his experience in teaching this course in the beginning. Andy Weiner is the Person in Charge (PIC) of this course at Purdue University, but he has generously and freely let me teach this course without interference. Mark Lundstrom gave me useful feedback on Chapters 38 and 39 on the quantum theory of light. I like to thank Dan Jiao for sharing her recent stellar contributions to fast algorithms in computational electromagnetics. Thanks also to Andy Weiner and Mahdi Hosseini for sharing their fascinating advances in quantum optics from their research group. I like to thank Erhan Kudeki of Illinois who always takes an active interest on my writing on this subject matter.

Also, I am thankful to Dong-Yeop Na for helping teach part of this course as well as collaborating on making important foray and advances in quantum electromagnetics when he was at Purdue. Thanks also are due the other members of our research team, Boyuan Zhang, Jie Zhu, Ivan Okhmatovskii, Akila Murugesan, and Sina Vaezi for supporting this course and also to Robert Hsueh-Yung Chao who took time to read the lecture notes and gave me some very useful feedback. Recently, Chris Ryu and Thomas Roth also gave useful feedback on the last two chapters of these notes.

Thomas Roth, Dong-Yeop Na, and I recently taught short courses on quantum electromagnetics at AP-S/URSI, Singapore 2021, Denver 2022, and Portland 2023. Some of the materials for the short courses are factored into the last two chapters of the lecture notes. We have also collaborated on research in this exciting emerging topic.

Recently, Dezhi Wang, Zekui Jia, and Chris Ryu helped proofread these lecture notes, making the teaching of this subject matter simpler. Special thanks go to Dezhi Wang for suggestion on how to make Chapter 38 easier to read.

I am also indebted to my wife, Chew-Chin Phua, for her life-long support, patience, and understanding, especially when I was putting together these lecture notes. We also have to make adjustments by moving from U of Illinois, Urbana-Champaign to Purdue U, West Lafayette.

Contents

[1]This section is mathematically complicated. It can be skipped on first reading.
[2]The derivation in this section is complex, but worth the pain, since this knowledge was not discovered until the 1960s.
[3]May be skipped on first reading.

[4]This section can be skipped on first reading.

[5]These polarizations are also variously know as TE_z, or the s and p polarizations, a descendent from the notations for acoustic waves where s and p stand for shear and pressure waves, respectively.

[6]Also known as TM_z polarization.

Part I

Fundamentals, Complex Media, Theorems and Principles

Chapter 1

Introduction, Maxwell's Equations

In the beginning, this field is either known as the field of electricity and magnetism or the field of optics. But later, as we shall discuss, these two fields are found to be based on the same set equations, the Maxwell's equations. Maxwell's equations unified these two fields, and it is common to call the study of electromagnetic theory based on Maxwell's equations *electromagnetics.* It has wide-ranging applications from statics to ultra-violet light in the present world with impact on many different technologies.

1.1 Importance of Electromagnetics

We will explain why electromagnetics is so important, and its impact on very many different areas. Then we will give a brief history of electromagnetics, and how it has evolved in the modern world. Next we will go briefly over Maxwell's equations in their full glory. But we will begin the study of electromagnetics by focussing on static problems which are valid in the long-wavelength limit or at zero frequency.

Electromagnetics has been based on Maxwell's equations, which are the result of the seminal work of James Clerk Maxwell completed in 1865, after his presentation to the British Royal Society in 1864. He was very much inspired by the experimentally motivated Coulomb's law and Gauss's law (1785), Ampere's law (1823), Faraday's law (1835). It has been over 150 years ago now, and this is a long time compared to the leaps and bounds progress we have made in technological advancements. Nevertheless, research in electromagnetics has continued unabated despite its age. The reason is that electromagnetics is extremely useful as it is pervasive, and has impacted a large sector of modern technologies.

To understand why electromagnetics is so useful, we have to understand first a few points about Maxwell's equations.

- Maxwell's equations are valid over a vast length scale from subatomic dimensions to galactic dimensions. Hence, these equations are valid over a vast range of wavelengths: from static to ultra-violet wavelengths.[1]

[1] Current lithography process is working with using deep ultra-violet light with a wavelength of 193 nm and

- Maxwell's equations are relativistic invariant in the parlance of special relativity [1]. In fact, Einstein was motivated with the theory of special relativity in 1905 by Maxwell's equations [2]. These equations look the same, irrespective of what inertial reference frame[2] one is in.

- Maxwell's equations are valid in the quantum regime, as it was demonstrated by Paul Dirac in 1927 [3]. Hence, many methods of calculating the response of a medium to classical field can be applied in the quantum regime also. When electromagnetic theory is combined with quantum theory, the field of quantum optics came about. Roy Glauber won a Nobel prize in 2005 because of his work in this area [4]. Alain Aspect, John F. Clausser, and Anton Zeilinger won the Nobel prize in 2022 for their works in quantum optics.

- Maxwell's equations and the pertinent gauge theory has inspired Yang-Mills theory (1954) [5], which is also known as a generalized electromagnetic theory. Yang-Mills theory is motivated by differential forms in differential geometry [6]. To quote from Misner, Thorne, and Wheeler, "Differential forms illuminate electromagnetic theory, and electromagnetic theory illuminates differential forms." [7, 8]

- Maxwell's equations are some of the most accurate physical equations that have been validated by experiments. In 1985, Richard Feynman wrote that electromagnetic theory had been validated to one part in a billion.[3] Now, it has been validated to one part in a trillion (Aoyama et al, Styer, 2012).[4]

- As a consequence, electromagnetics has permeated many technologies, and has a tremendous impact in science and technology. This is manifested in electrical engineering, optics, wireless and optical communications, computers, remote sensing, subsurface sensing, bio-medical engineering etc. It is expected that quantum electromagnetics (the quantum extension of electromagnetics) will grow in importance as quantum technologies develop.

---

subsequently, 13.5 nm.

[2] An inertial reference frame is a coordinate frame that is traveling at a velocity v.

[3] This means that if a jet is to fly from New York to Los Angeles, an error of one part in a billion means an error of a few millmeters.

[4] This means an error of a hairline, if one were to shoot a light beam from the earth to the moon.

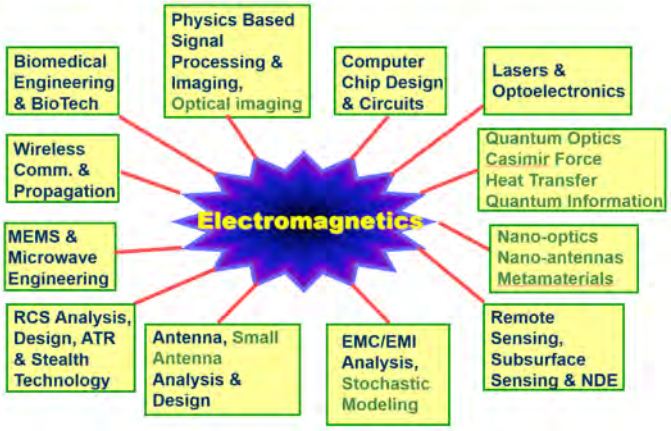


Figure 1.1: The impact of electromagnetics in many technologies. The areas in blue are prevalent areas impacted by electromagnetics some 20 years ago [9], and the areas in brown are modern emerging areas impacted by electromagnetics.

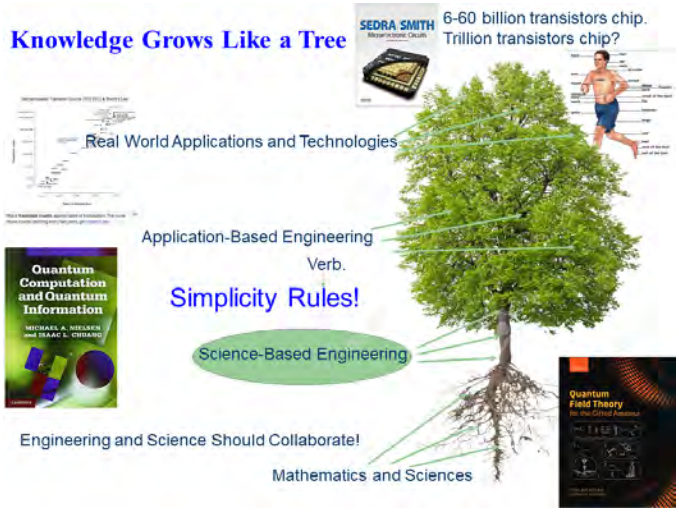


Figure 1.2: Knowledge grows like a tree. Engineering knowledge and real-world applications are driven by fundamental knowledge from math and the sciences. At a university, we do science-based engineering research that can impact wide-ranging real-world applications. But everyone is equally important in transforming our society. Just like the parts of the human body: no one can claim that one is more important than the others (1 Corinthians 12).

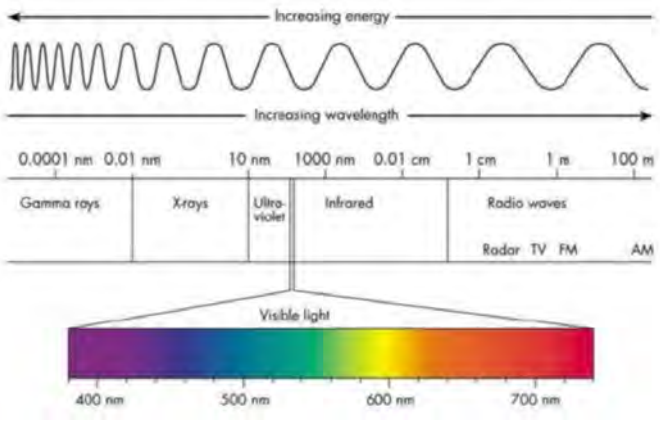


Figure 1.3: The electromagnetics spectrum goes from statics to UV. Deep UV and extreme UV (EUV) are now used in nano-lithography, while statics is used in circuit theory and geophysical exploration.

Figure 1.2 shows how knowledge are driven by basic math and science knowledge. Its growth is like a tree. Due to the vast ocean of knowledge that we are immersed in, it is important that we collaborate, especially between engineering and science, to develop technologies that can transform this world.

1.1.1 The Electromagnetic Spectrum

The electromagnetic field has been used from very low frequencies to very high frequencies. At very low frequencies, ultra-low frequency (ULF) <3 Hz, extremely-low frequency (ELF) 3-3000 Hz, very low frequency (VLF) 3 KHz to 30 KHz have been used to probe the earth surface, and submarine communication because of their deeper penetration depths. The AM radio station operating in the several 100 KHz has wavelength of several 100 m. FM radio are in the 100 MHz range, while TV stations operate in the several 100 MHz. Microwaves have wavelength of order of cm, and infra-red light ranges from 1000 μm to 1 μm. The visible spectrum ranges from 700 nm to 400 nm. Ultra-violet (UV) light ranges from 400 nm to 100 nm, while X-ray is generally below 100 nm to 1 nm. Gamma ray is generally below 1 nm.

UV light of 193 nm were used for nano-lithography while extreme UV (EUV) of 13.5 nm are now also used. X-ray is important for imaging, while gamma ray is used for some medical applications. The lights above UV are harmful to the human body.

1.1.2 A Brief History of Electromagnetics

Electricity and magnetism have been known to mankind for a long time. Also, the physical properties of light have been known. But the field of electricity and magnetism, now termed electromagnetics in the modern world, has been thought to be governed by different physical laws

as opposed to those for optics. This is understandable as the physics of electricity and magnetism is quite different of the physics of optics as they were known to humans then.

For example, lode stone was known to the ancient Greek and Chinese around 600 BC to 400 BC. Compass was used in China since 200 BC. Static electricity was reported by the Greek as early as 400 BC. But these curiosities did not make an impact until the age of telegraphy. The coming about of telegraphy was due to the invention of the voltaic cell or the galvanic cell in the late 1700's, by Luigi Galvani and Alesandro Volta [10]. It was soon discovered that two pieces of wire, connected to a voltaic cell, can transmit information at a distance.

So by the early 1800's this possibility had spurred the development of telegraphy. Both André-Marie Ampére (1823) [11, 12] and Michael Faraday (1838) [13] did experiments to better understand the properties of electricity and magnetism. And hence, Ampere's law and Faraday law were named after them. Kirchhoff voltage and current laws were also developed in 1845 to help better understand telegraphy [14, 15]. Despite these laws, the technology of telegraphy was poorly understood. For instance, it was not known as to why the telegraphy signal was distorted. Ideally, the signal should be a digital signal switching between one's and zero's, but the digital signal lost its shape rapidly along a telegraphy line.[5]

It was not until 1865 that James Clerk Maxwell [17] put in the missing term in Ampere's law, the displacement current term, only then the mathematical theory for electricity and magnetism was complete. Ampere's law is now known as generalized Ampere's law. The complete set of equations are now named Maxwell's equations in honor of James Clerk Maxwell.[6]

The rousing success of Maxwell's theory was that it predicted wave phenomena, as they have been observed along telegraphy lines. But it was not until 23 years later that Heinrich Hertz in 1888 [19] did experiments to prove that electromagnetic field can propagate through space across a room. This illustrates the difficulty of slow knowledge dissemination then when new knowledge was discovered. Moreover, from experimental measurement of the permittivity and permeability of matter, it was decided that electromagnetic wave moves at a tremendous speed. But the velocity of light has been known for a long while from astronomical observations (Roemer, 1676) [20]. The interference phenomena in light has been observed in Newton's ring (1704) [21]. When these pieces of information were combined together, it was decided that electricity and magnetism, and optics, are actually governed by the same physical law or Maxwell's equations. And optics and electromagnetics are unified into one field!

[5] As a side note, in 1837, Morse invented the Morse code for telegraphy [16]. There were cross pollination of ideas across the Atlantic ocean despite the distance. In fact, Benjamin Franklin associated lightning with electricity in the latter part of the 18-th century. Also, notice that electrical machinery was invented in 1832 even though electromagnetic theory was not fully understood.

[6] However, it was Oliver Heaviside (1850-1925) who distilled Maxwell's equations into four equations that are found in electromagnetics textbooks now [18].

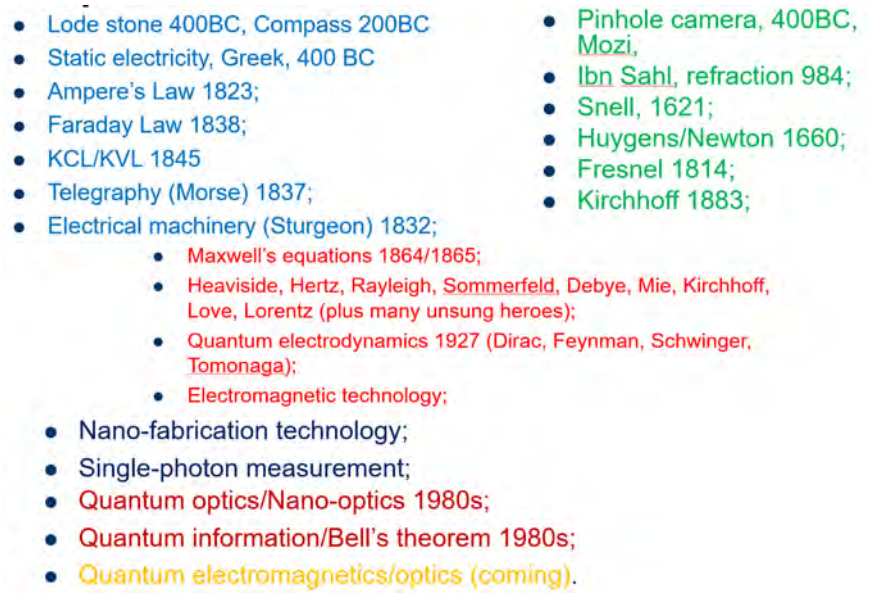


Figure 1.4: A brief history of electromagnetics and optics as depicted in this figure. In the early days, it was thought that optics is a different discipline from electricity and magnetism. Then after 1865, the two fields are unified and governed by Maxwell's equations.

In Figure 1.4, a brief history of electromagnetics and optics is depicted. In the beginning, it was thought that electricity and magnetism, and optics were governed by different physical laws. Low frequency electromagnetics was governed by the understanding of fields and their interaction with media. Optical phenomena were governed by ray optics, reflection and refraction of light. But the advent of Maxwell's equations in 1865 revealed that they can be unified under electromagnetic theory. Then solving Maxwell's equations becomes a rewarding mathematical endeavor.

The photoelectric effect [22, 23], and Planck radiation law [24] point to the fact that electromagnetic energy is manifested in terms of packets of energy, indicating the corpuscular nature of light. Each unit of this energy is now known as the photon. A photon carries an energy packet equal to $\hbar\omega$, where ω is the angular frequency of the photon and the Planck constant $\hbar = 6.626 \times 10^{-34}$ J s, which is a very small constant. Hence, the higher the frequency, the easier it is to detect this packet of energy, or feel the graininess of electromagnetic energy. Eventually, in 1927 [3], quantum theory was incorporated into electromagnetics, and the quantum nature of light gives rise to the field of quantum optics. Recently, even microwave photons have been measured [25, 26]. They are difficult to detect because of the low frequency of microwave (10^9 Hz) compared to optics (10^{15} Hz): a microwave photon carries a packet of energy about a million times smaller than that of an optical photon.

The progress in nano-fabrication [27] allows one to make optical components that are sub-

optical wavelength as the wavelength of blue light is about 450 nm.[7] As a result, interaction of light with nano-scale optical components requires the solution of Maxwell's equations in its full glory, whereas traditionally, ray optics were used to describe many optical phenomena.

In the early days of quantum theory, there were two prevailing theories of quantum interpretation. Quantum measurements were found to be random. In order to explain the probabilistic nature of quantum measurements, Einstein posited that a random hidden variable caused the random outcome of an experiment. On the other hand, the Copenhagen school of interpretation led by Niels Bohr, asserted that the outcome of a quantum measurement is not known until after a measurement [28].

In 1960s, Bell's theorem (by John Steward Bell) [29] said that an inequality should be satisfied if Einstein's hidden variable theory was correct. Otherwise, the violation of the inequality implies that the Copenhagen school of interpretation should prevail. However, experimental measurement showed that the inequality was violated, favoring the Copenhagen school of quantum interpretation [28]. This interpretation says that a quantum state is in a linear superposition of states before a measurement. But after a measurement, a quantum state "collapses" to the state that is measured. This implies that quantum information can be hidden *incognito* in a quantum state. Hence, for a quantum particle, such as a photon, its quantum state is unknown until after its measurement. In other words, quantum theory is "spooky" or "weird". This also has the profound and beautiful implication that "our karma is not written on our forehead when we were born, our future is in our own hands!". This leads to growing interest in quantum information and quantum communication using photons. Quantum technology with the use of photons, an electromagnetic quantum particle, is a subject of growing interest.

1.2 Maxwell's Equations in Integral Form

Even though experimentally motivated, Maxwell's equations can be presented as fundamental postulates.[8] We will present them in their integral forms, but will not belabor them until later.

$$\oint_C \mathbf{E}\cdot\boldsymbol{dl} = -\frac{d}{dt}\iint_S \mathbf{B}\cdot\boldsymbol{dS} \qquad \text{Faraday's Law (1838)} \tag{1.2.1}$$

$$\oint_C \mathbf{H}\cdot\boldsymbol{dl} = \frac{d}{dt}\iint_S \mathbf{D}\cdot\boldsymbol{dS} + I \qquad \text{Ampere's Law (1823)} \tag{1.2.2}$$

$$\oiint_S \mathbf{D}\cdot\boldsymbol{dS} = Q \qquad \text{Gauss's or Coulomb's Law (1785)} \tag{1.2.3}$$

$$\oiint_S \mathbf{B}\cdot\boldsymbol{dS} = 0 \qquad \text{Gauss's Law (1835)} \tag{1.2.4}$$

The units of the basic quantities above are given as:

$$\mathbf{E}\text{: V/m} \qquad\qquad \mathbf{H}\text{: A/m}$$

[7] Size of the smallest transistor now is about 5 nm, while the size of the coronovirus is about 50 to 140 nm.

[8] Postulates in physics are similar to axioms in mathematics. They are assumptions that need not be proved.

$\mathbf{D}$: C/m^2	$\mathbf{B}$: W/m^2
I: A	Q: C

where V=volts, A=amperes, C=coulombs, and W=webers.

In this course, we use a boldface to denote a vector, and a hat to denote a unit vector. Hence, a vector can be written as $\mathbf{E} = \hat{x}E_x + \hat{y}E_y + \hat{z}E_z$. where $\hat{x}$, $\hat{y}$, and $\hat{z}$ are unit vectors in Cartesian coordinates. In some books, alternatively, a vector is written as $\mathbf{E} = (E_x, E_y, E_z)$.

Before we close this section, it is to be noted that (1.2.3) and (1.2.4) are derivable (1.2.1) and (1.2.2) for time-varying problems. Therefore, for time-varying problems, not all four Maxwell's equations are independent of each other. They are only independent of each other for electrostatic problems. We will verify this fact later.

1.3 Static Electromagnetics

In statics, the field is assumed to be non-time-varying. Hence all the time dependence terms can be removed from Maxwell's equations, and we have

$$\oint_C \mathbf{E} \cdot \boldsymbol{dl} = 0 \qquad \text{Faraday's Law} \tag{1.3.1}$$

$$\oint_C \mathbf{H} \cdot \boldsymbol{dl} = I \qquad \text{Ampere's Law} \tag{1.3.2}$$

$$\oiint_S \mathbf{D} \cdot \boldsymbol{dS} = Q \qquad \text{Gauss's or Coulomb's Law} \tag{1.3.3}$$

$$\oiint_S \mathbf{B} \cdot \boldsymbol{dS} = 0 \qquad \text{Gauss's Law} \tag{1.3.4}$$

The first equation above, which is the static form of Faraday's law also gives rise to Kirchhoff voltage law. The second equation is the original form of Ampere's law where displacement current was ignored. The third and the fourth equations remain unchanged compared to the time-varying (dynamic) form of Maxwell's equations.

1.3.1 Coulomb's Law (Statics)

This law, developed in 1785 [30], expresses the force between two charges q_1 and q_2. If these charges are positive, the force is repulsive and it is given by

$$\mathbf{f}_{1\to 2} = \frac{q_1 q_2}{4\pi\varepsilon r^2}\hat{\mathbf{r}}_{12} \tag{1.3.5}$$

where the units are: $\mathbf{f}$ (force): newton
q (charge): coulomb
ε (permittivity): farad/meter
r (distance between q_1 and q_2): meter

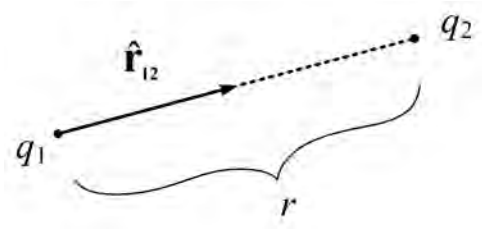


Figure 1.5: The force between two charges q_1 and q_2. The force is repulsive if the two charges have the same sign.

$\hat{\mathbf{r}}_{12}$= unit vector pointing from charge 1 to charge 2

$$\hat{\mathbf{r}}_{12} = \frac{\mathbf{r}_2 - \mathbf{r}_1}{|\mathbf{r}_2 - \mathbf{r}_1|}, \qquad r = |\mathbf{r}_2 - \mathbf{r}_1| \tag{1.3.6}$$

Using the definition for unit vector, the force between two charges can also be rewritten as

$$\mathbf{f}_{1\to 2} = \frac{q_1 q_2 (\mathbf{r}_2 - \mathbf{r}_1)}{4\pi\varepsilon |\mathbf{r}_2 - \mathbf{r}_1|^3}, \qquad (\mathbf{r}_1, \mathbf{r}_2 \text{ are position vectors}) \tag{1.3.7}$$

1.3.2 Electric Field (Statics)

The electric field $\mathbf{E}$ is defined as the force per unit charge [31]. For two charges, one of charge q and the other one of incremental charge Δq, the force between the two charges, according to Coulomb's law (1.3.5), is

$$\mathbf{f} = \frac{q\Delta q}{4\pi\varepsilon r^2}\hat{\mathbf{r}} \tag{1.3.8}$$

where $\hat{\mathbf{r}}$ is a unit vector pointing from charge q to the incremental charge Δq. Then the electric field $\mathbf{E}$, which is the force per unit charge, is given by

$$\mathbf{E} = \frac{\mathbf{f}}{\triangle q} = \frac{q}{4\pi\varepsilon r^2}\hat{\mathbf{r}}, \qquad (\text{V/m}) \tag{1.3.9}$$

Therefore, in general, the electric field $\mathbf{E}(\mathbf{r})$ at location $\mathbf{r}$ from a point charge q at $\mathbf{r}'$, using the definition for $\hat{\mathbf{r}}$, is given by

$$\mathbf{E}(\mathbf{r}) = \frac{q(\mathbf{r} - \mathbf{r}')}{4\pi\varepsilon |\mathbf{r} - \mathbf{r}'|^3} \tag{1.3.10}$$

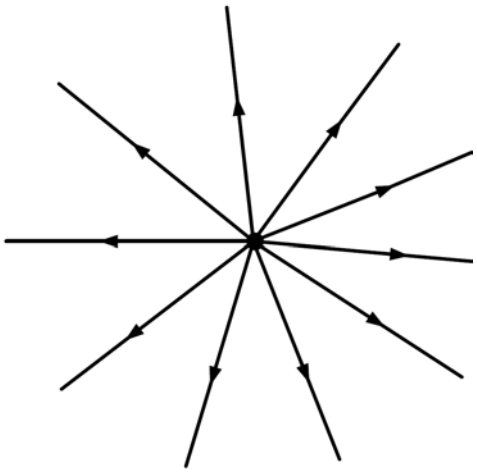


Figure 1.6: Emanating **E** field from an electric point charge as depicted by (1.3.9) and (1.3.10). The physical meaning is that for a positive charge q

If one knows **E** due to a point charge, one will know **E** due to any charge distribution because any charge distribution can be decomposed into sum of point charges. For instance, if there are N point charges each with amplitude q_i, then by the principle of linear superposition assuming that linearity holds, the total field produced by these N charges is

$$\mathbf{E}(\mathbf{r}) = \sum_{i=1}^{N} \frac{q_i(\mathbf{r} - \mathbf{r}_i)}{4\pi\varepsilon|\mathbf{r} - \mathbf{r}_i|^3} \tag{1.3.11}$$

where $q_i = \varrho(\mathbf{r}_i)\Delta V_i$ is the incremental charge at $\mathbf{r}_i$ enclosed in the volume ΔV_i. In the continuum limit, the above becomes

$$\mathbf{E}(\mathbf{r}) = \int_V \frac{\varrho(\mathbf{r}')(\mathbf{r} - \mathbf{r}')}{4\pi\varepsilon|\mathbf{r} - \mathbf{r}'|^3} dV \tag{1.3.12}$$

In other words, the total field, by the principle of linear superposition, is the integral summation of the contributions from the distributed charge density $\varrho(\mathbf{r})$.

1.3.3 Gauss's Law for Electric Flux (Statics)

This law is also known as Coulomb's law as they are closely related to each other. Apparently, this simple law was first expressed by Joseph Louis Lagrange in 1773 [32] and later, reexpressed by Gauss in 1813 (Wikipedia).

This law can be expressed as

$$\oiint_S \mathbf{D} \cdot d\mathbf{S} = Q \tag{1.3.13}$$

where $\mathbf{D}$ is electric flux density with unit C/m^2 and $\mathbf{D} = \varepsilon \mathbf{E}$, $d\mathbf{S}$ is an incremental surface at the point on S given by $dS\hat{\mathbf{n}}$ where $\hat{\mathbf{n}}$ is the unit normal pointing outward away from the surface, and Q is total charge enclosed by the surface S.

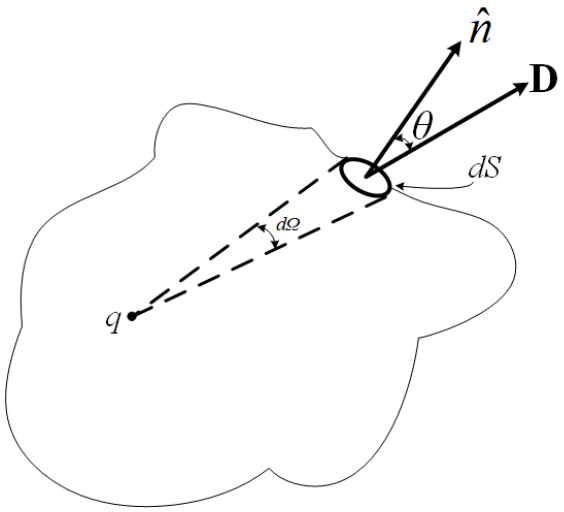


Figure 1.7: Electric flux through an incremental surface dS where $\hat{n}$ is the unit normal, and $\mathbf{D}$ is the electric flux density passing through the incremental surface.

The left-hand side of (1.3.13) represents a surface integral over a closed surface S. To understand it, one can break the surface into a sum of incremental surfaces ΔS_i, with a local unit normal $\hat{\mathbf{n}}_i$ associated with it. The surface integral can then be approximated by a summation

$$\oiint_S \mathbf{D} \cdot d\mathbf{S} \approx \sum_i \mathbf{D}_i \cdot \hat{\mathbf{n}}_i \Delta S_i = \sum_i \mathbf{D}_i \cdot \Delta \mathbf{S}_i \tag{1.3.14}$$

where one has defined the incremental surface $\Delta \mathbf{S}_i = \hat{\mathbf{n}}_i \Delta S_i$. In the limit when ΔS_i becomes infinitesimally small, the summation becomes a surface integral.

1.3.4 Derivation of Gauss's Law from Coulomb's Law (Statics)

From Coulomb's law, the ensuing electric field due to a point charge, and the electric flux is

$$\mathbf{D} = \varepsilon \mathbf{E} = \frac{q}{4\pi r^2}\hat{\mathbf{r}} \tag{1.3.15}$$

When a closed spherical surface S is drawn around the point charge q, by symmetry, the electric flux though every point of the surface is the same. Moreover, the normal vector $\hat{\mathbf{n}}$ on the surface is just $\hat{\mathbf{r}}$. Consequently, $\mathbf{D}\cdot\hat{\mathbf{n}} = \mathbf{D}\cdot\hat{\mathbf{r}} = q/(4\pi r^2)$, which is a constant on a spherical of radius r. Hence, we conclude that for a point charge q, and the pertinent electric flux $\mathbf{D}$ that it produces on a spherical surface satisfies,

$$\oiint_S \mathbf{D}\cdot d\mathbf{S} = 4\pi r^2 \mathbf{D}\cdot\hat{\mathbf{n}} = 4\pi r^2 D_r = q \tag{1.3.16}$$

Therefore, Gauss's law is satisfied by a point charge.

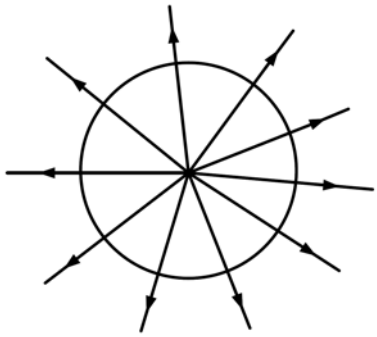


Figure 1.8: Electric flux from a point charge satisfies Gauss's law.

Even when the shape of the spherical surface S is distorted from a sphere to an arbitrary shape surface S, it can be shown that the total flux through S is still q. In other words, the total flux through sufaces S_1 and S_2 in Figure 1.9 are the same.

This can be appreciated by taking a sliver of the angular sector as shown in Figure 1.10. Here, ΔS_1 and ΔS_2 are two incremental surfaces intercepted by this sliver of angular sector. The amount of flux passing through this incremental surface is given by $d\mathbf{S}\cdot\mathbf{D} = \hat{\mathbf{n}}\cdot\mathbf{D}\Delta S = \hat{\mathbf{n}}\cdot\hat{\mathbf{r}}D_r\Delta S$. Here, $\mathbf{D} = \hat{\mathbf{r}}D_r$ is pointing in the $\hat{\mathbf{r}}$ direction. In ΔS_1, $\hat{n}$ is pointing in the $\hat{\mathbf{r}}$ direction. But in ΔS_2, the incremental area has been enlarged by that $\hat{\mathbf{n}}$ not aligned with $\mathbf{D}$. But this enlargement is compensated by $\hat{\mathbf{n}}\cdot\hat{\mathbf{r}}$. Also, ΔS_2 has grown bigger, but the flux at ΔS_2 has grown smaller by the ratio of $(r_2/r_1)^2$. Finally, the two fluxes are equal in the limit that the sliver of angular sector becomes infinitesimally small. This proves the assertion that the total fluxes through S_1 and S_2 are equal. Since the total flux from a point charge q through a closed surface is independent of its shape, but always equal to q, then if we have a total charge Q which can be expressed as the sum of point charges, namely.

$$Q = \sum_i q_i \tag{1.3.17}$$

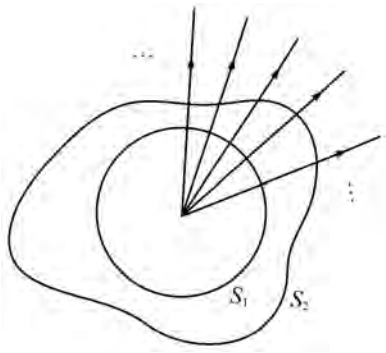


Figure 1.9: Same amount of electric flux from a point charge passes through two surfaces S_1 and S_2. This allows Gauss's law for electric flux to be derivable from Coulomb's law for statics.

Then the total flux through a closed surface equals the total charge enclosed by it, which is the statement of Gauss's law or Coulomb's law.

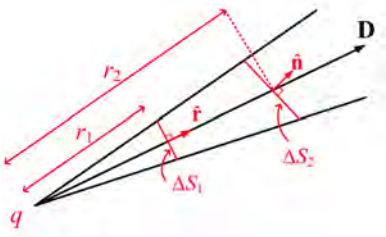


Figure 1.10: When a sliver of angular sector is taken, same amount of electric flux from a point charge passes through two incremental surfaces ΔS_1 and ΔS_2 at different distances from the point charge.

Exercises for Lecture 1

Problem 1-1:

(i) Explain why the electric flux going through ΔS_1 and ΔS_2 are the same in Figure 1.10. To make it simpler, you can explain for the 2D case.

(ii) Find the field due to a ring of charges with line charge density ϱ C/m as shown in the figure (courtesy of Ramo, Whinnery, and Van Duzer). **Hint:** Use symmetry.

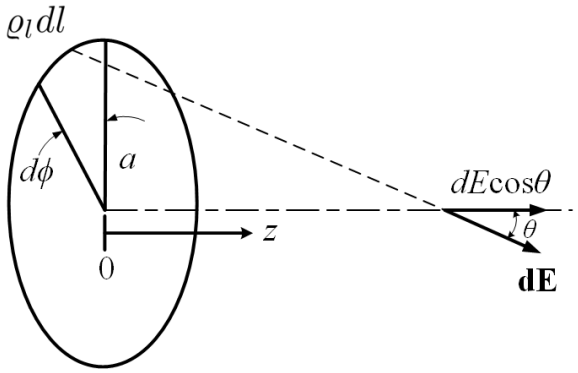


Figure 1.11: Electric field of a ring of charge (courtesy of Ramo, Whinnery, and Van Duzer [33]).

(iii) What is the electric field between coaxial cylinders with surface charge densities of ϱ_s and $-\varrho_s$ of unit length in a coaxial cable? **Hint:** Use symmetry and cylindrical coordinates to

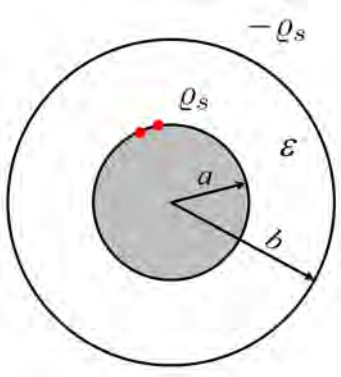


Figure 1.12: Figure for Problem 1-1 for a coaxial cylinder.

express $\mathbf{E} = \hat{\rho} E_\rho$ and apply Gauss's law.

(iv) The figure shows a sphere of uniform volume charge density ϱ. Find the electric field $\mathbf{E}$ as a function of distance r from the center of the sphere. **Hint:** Again, use symmetry and spherical coordinates to express $\mathbf{E} = \hat{r}E_r$ and apply Gauss's law.

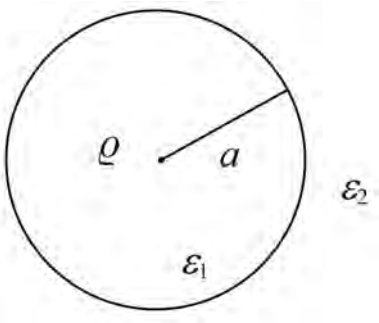


Figure 1.13: Figure for Problem 1-1 for a sphere with uniform volume charge density ϱ.

(v) Given an infinitely long cylindrical circular wire carrying a DC current I, find the magnetic field around the wire using symmetry argument, and Ampere's law.

Chapter 2

Maxwell's Equations, Differential Operator Form

Maxwell's equations were originally written in integral form as has been shown in the previous lecture. Integral forms have nice physical meaning and can be easily related to experimental measurements. However, the differential operator form[1] can be easily converted to differential equations or partial differential equations where a whole sleuth of mathematical methods and numerical methods can be deployed. Therefore, it is prudent to derive the differential operator form of Maxwell's equations.

2.1 Gauss's Divergence Theorem

We will first prove Gauss's divergence theorem.[2] The divergence theorem is one of the most important theorems in vector calculus [33, 34, 35, 36]. It says that:

$$\iiint_V dV \nabla \cdot \mathbf{D} = \oiint_S \mathbf{D} \cdot d\mathbf{S} \tag{2.1.1}$$

The right-hand side of the above is the total electric flux $\mathbf{D}$ that comes out of the surface S. In the above, assuming that $V \to \Delta V$, an infinitesimal volume, then $\nabla \cdot \mathbf{D}$ is defined as

$$\nabla \cdot \mathbf{D} = \lim_{\Delta V \to 0} \frac{\oiint_{\Delta S} \mathbf{D} \cdot d\mathbf{S}}{\Delta V} \tag{2.1.2}$$

The above can be used to derive the explicit for of $\nabla \cdot \mathbf{D}$ as shall be shown later. The above implies that the divergence of the electric flux $\mathbf{D}$, or $\nabla \cdot \mathbf{D}$ is given by first computing the flux coming (or oozing) out of a small volume ΔV surrounded by a small surface ΔS and taking their ratio

[1] We caution ourselves not to use the term "differential forms" which has a different meaning used in differential geometry for another form of Maxwell's equations.

[2] Named after Carl Friedrich Gauss, a precocious genius who lived between 1777-1855.

as shown on the right-hand side of the above. As shall be shown, the ratio has a finite limit and eventually, we will find a simplified expression for it. We know that if $\Delta V \approx 0$ or small, then the above implies that,

$$\Delta V \nabla \cdot \mathbf{D} \approx \oiint_{\Delta S} \mathbf{D} \cdot d\mathbf{S} \tag{2.1.3}$$

First, we assume that a volume V has been discretized[3] into a sum of small cuboids, where the i-th cuboid has a volume of ΔV_i as shown in Figure 2.1. Then

$$V \approx \sum_{i=1}^{N} \Delta V_i \tag{2.1.4}$$

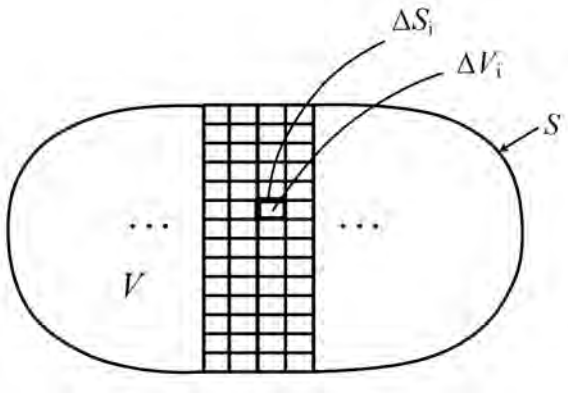


Figure 2.1: The discretization of a volume V into a sum of small volumes ΔV_i each of which is a small cuboid. Stair-casing error occurs near the boundary of the volume V but the error diminishes as $\Delta V_i \to 0$.

[3]Other terms used are "tesselated", "meshed", or "gridded".

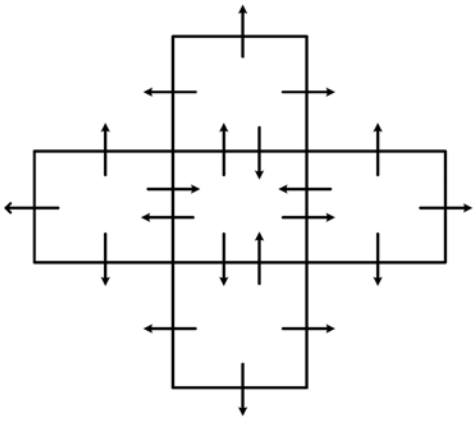


Figure 2.2: Fluxes from adjacent cuboids cancel each other leaving only the fluxes at the boundary that remain uncancelled. Please imagine that there is a third dimension of the cuboids in this picture where it comes out of the paper.

Then from (2.1.2) and (2.1.3), for the i-th cuboid,

$$\Delta V_i \nabla \cdot \mathbf{D}_i \approx \oiint_{\Delta S_i} \mathbf{D}_i \cdot d\mathbf{S}_i \tag{2.1.5}$$

By summing the above over all the cuboids, or over i, one gets

$$\sum_i \Delta V_i \nabla \cdot \mathbf{D}_i \approx \sum_i \oiint_{\Delta S_i} \mathbf{D}_i \cdot d\mathbf{S}_i \approx \oiint_S \mathbf{D} \cdot d\mathbf{S} \tag{2.1.6}$$

The last approximation follows, because it is easily seen that the fluxes out of the inner surfaces of the cuboids cancel each other, leaving only fluxes flowing out of the cuboids at the edge of the volume V as explained in Figure 2.2. The right-hand side of the above equation (2.1.6) becomes a surface integral over the surface S except for the stair-casing approximation (see Figure 2.1). However, this approximation becomes increasingly good as $\Delta V_i \to 0$. Moreover, the left-hand side of (2.1.6) becomes a volume integral, and we have

$$\iiint_V dV \nabla \cdot \mathbf{D} = \oiint_S \mathbf{D} \cdot d\mathbf{S} \tag{2.1.7}$$

The above is the well known Gauss's divergence theorem.

2.1.1 Some Details

Next, we will derive the details of the definition embodied in (2.1.2). To this end, we evaluate the numerator of the right-hand side carefully, in accordance to Figure 2.3.

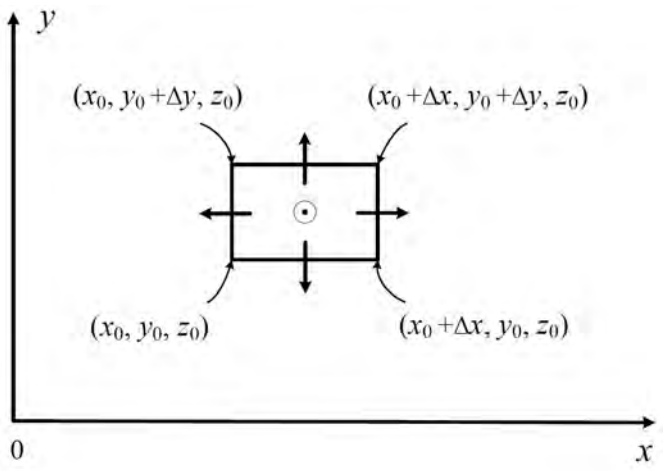


Figure 2.3: Figure to illustrate the calculation of fluxes from a small cuboid where a corner of the cuboid is located at (x_0, y_0, z_0). There is a third z dimension of the cuboid not shown, and coming out of the paper. Hence, this cuboid, unlike that shown in the figure, has six faces.

Accounting for the fluxes going through all the six faces, assigning the appropriate signs in accordance with the fluxes leaving and entering the cuboid, one arrives at the following six terms

$$\begin{aligned}\oiint_{\Delta S} \mathbf{D}\cdot d\mathbf{S} \approx & -\mathrm{D}_x(x_0,y_0,z_0)\Delta y\Delta z + \mathrm{D}_x(x_0+\Delta x,y_0,z_0)\Delta y\Delta z \\ & -\mathrm{D}_y(x_0,y_0,z_0)\Delta x\Delta z + \mathrm{D}_y(x_0,y_0+\Delta y,z_0)\Delta x\Delta z \\ & -\mathrm{D}_z(x_0,y_0,z_0)\Delta x\Delta y + \mathrm{D}_z(x_0,y_0,z_0+\Delta z)\Delta x\Delta y \end{aligned} \tag{2.1.8}$$

Factoring out the volume of the cuboid $\Delta V = \Delta x\Delta y\Delta z$ in the above, one gets

$$\begin{aligned}\oiint_{\Delta S} \mathbf{D}\cdot d\mathbf{S} \approx \Delta V \{ & [D_x(x_0+\Delta x,\ldots) - D_x(x_0,\ldots)]/\Delta x \\ & + [D_y(\ldots,y_0+\Delta y,\ldots) - D_y(\ldots,y_0,\ldots)]/\Delta y \\ & + [D_z(\ldots,z_0+\Delta z) - D_z(\ldots,z_0)]/\Delta z\} \end{aligned} \tag{2.1.9}$$

Or that

$$\frac{\oiint \mathbf{D}\cdot d\mathbf{S}}{\Delta V} \approx \frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z} \tag{2.1.10}$$

In the limit when $\Delta V \to 0$, then

$$\lim_{\Delta V\to 0} \frac{\oiint \mathbf{D}\cdot d\mathbf{S}}{\Delta V} = \frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z} = \nabla\cdot\mathbf{D} \tag{2.1.11}$$

where (2.1.2) has been used for the above definition of $\nabla \cdot \mathbf{D}$. Furthermore,

$$\nabla = \hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z} \tag{2.1.12}$$

$$\mathbf{D} = \hat{x}D_x + \hat{y}D_y + \hat{z}D_z \tag{2.1.13}$$

The above is the definition of the divergence operator in Cartesian coordinates. The divergence operator $\nabla\cdot$ has its complicated representations in cylindrical and spherical coordinates, a subject that we would not delve into in this course. But they can be derived, and are best looked up at the back of some textbooks on electromagnetics [36].

Consequently, one obtains the Gauss's divergence theorem given by

$$\iiint_V dV \nabla \cdot \mathbf{D} = \oiint_S \mathbf{D} \cdot d\mathbf{S} \tag{2.1.14}$$

2.1.2 Physical Meaning of Divergence Operator

The physical meaning of divergence is that if $\nabla \cdot \mathbf{D} \neq 0$ at a point in space, it implies that there are fluxes oozing or exuding from that point in space [37]. On the other hand, if $\nabla \cdot \mathbf{D} = 0$, it implies no net flux oozing out from that point in space. In other words, whatever flux that goes into the point must come out of it. The flux is then termed divergence free. Thus, $\nabla \cdot \mathbf{D}$ is a measure of how much sources or sinks exist for the flux at a point. The sum of these sources or sinks gives the amount of flux leaving or entering the surface that surrounds the sources or sinks.

Moreover, if one were to integrate a divergence-free flux over a volume V, and invoking Gauss's divergence theorem, one gets

$$\oiint_S \mathbf{D} \cdot d\mathbf{S} = 0 \tag{2.1.15}$$

In such a scenerio, whatever flux that enters the surface S must leave it. In other words, what comes in must go out of the volume V, or that flux is conserved. This is true of incompressible fluid flow [38], electric flux flow in a source free region, as well as magnetic flux flow, where the flux is conserved.

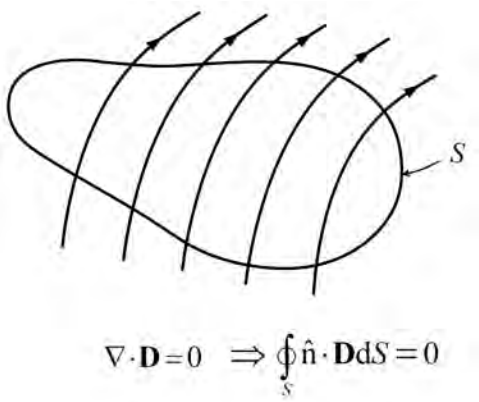


Figure 2.4: In an incompressible flux flow, flux is conserved: whatever flux that enters a volume V must leave the volume V.

2.1.3 Gauss's Law in Differential Operator Form

By further using Gauss's or Coulomb's law, it implies that

$$\oiint_S \mathbf{D} \cdot d\mathbf{S} = Q = \iiint dV \varrho \tag{2.1.16}$$

We can replace the left-hand side of the above by (2.1.14) to arrive at

$$\iiint_V dV \nabla \cdot \mathbf{D} = \iiint_V dV \varrho \tag{2.1.17}$$

When $V \to 0$, we arrive at the pointwise relationship, a relationship at an arbitrary point in space. Therefore,

$$\nabla \cdot \mathbf{D} = \varrho \tag{2.1.18}$$

2.2 Stokes's Theorem

The mathematical description of fluid flow was well established before the establishment of electromagnetic theory [39]. Hence, much mathematical description of electromagnetic theory uses the language of fluid. In mathematical notations, Stokes's theorem is[4]

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = \iint_S \nabla \times \mathbf{E} \cdot d\mathbf{S} \tag{2.2.1}$$

[4]Named after George Gabriel Stokes who lived between 1819 to 1903. In this course, we will use *Stokes'* and *Stokes's*, and its likes interchangeably.

In the above, the contour C is a closed contour, whereas the surface S is not closed.[5]

From the above definition, we can derive explicit form for $\nabla \times \mathbf{E}$. First, applying Stokes's theorem to a small surface ΔS, we define a curl operator[6] $\nabla\times$ at a point to be measured as

$$(\nabla \times \mathbf{E}) \cdot \hat{n} = \lim_{\Delta S \to 0} \frac{\oint_{\Delta C} \mathbf{E} \cdot d\mathbf{l}}{\Delta S} \tag{2.2.2}$$

In the above, $\mathbf{E}$ is a force per unit charge, and $\nabla \times \mathbf{E}$ is a vector. The above can be viewed as the definition of $\nabla \times \mathbf{E}$. Taking $\oint_{\Delta C} \mathbf{E} \cdot d\mathbf{l}$ as a measure of the torque or rotation of the field $\mathbf{E}$ around a small loop ΔC, the ratio of this rotation to the area of the loop ΔS has a limit when ΔS becomes infinitesimally small. This ratio is related to $(\nabla \times \mathbf{E}) \cdot \hat{n}$ where $\hat{n}$ is a unit normal to the surface ΔS. As in angular momentum, the direction of the torque is along the rotation axis of the force.

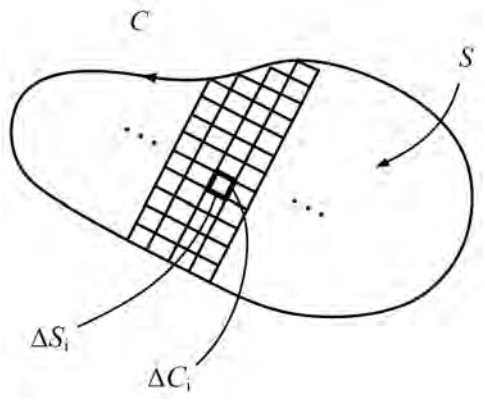


Figure 2.5: In proving Stokes's theorem, a closed contour C is assumed to enclose an open flat surface S. Then the surface S is tessellated into sum of small rects (rectangles) as shown. Stair-casing error at the boundary C vanishes in the limit when the rects are made vanishingly small.

First, the flat surface S enclosed by C is tessellated (also called meshed, gridded, or discretized) into sum of small rects (rectangles) as shown in Figure 2.5. Then, (2.2.2), is applied to one of these small rects to arrive at

$$\oint_{\Delta C_i} \mathbf{E}_i \cdot d\mathbf{l}_i = (\nabla \times \mathbf{E}_i) \cdot \Delta \mathbf{S}_i \tag{2.2.3}$$

where one defines $\Delta \mathbf{S}_i = \hat{n} \Delta S$. Next, we sum the above equation over i or over all the small rects

[5] In other words, C has no boundary whereas S has boundary. A closed surface S has no boundary like when we were proving Gauss's divergence theorem previously.

[6] Sometimes called a rotation operator.

to arrive at

$$\sum_i \oint_{\Delta C_i} \mathbf{E}_i \cdot d\mathbf{l}_i = \sum_i \nabla \times \mathbf{E}_i \cdot \Delta \mathbf{S}_i \tag{2.2.4}$$

Again, on the left-hand side of the above, all the contour integrals over the small rects cancel each other internal to S save for those on the boundary. In the limit when $\Delta S_i \to 0$, the left-hand side becomes a contour integral over the larger contour C, and the right-hand side becomes a surface integral over S. One arrives at Stokes's theorem, which is

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = \iint_S (\nabla \times \mathbf{E}) \cdot d\mathbf{S} \tag{2.2.5}$$

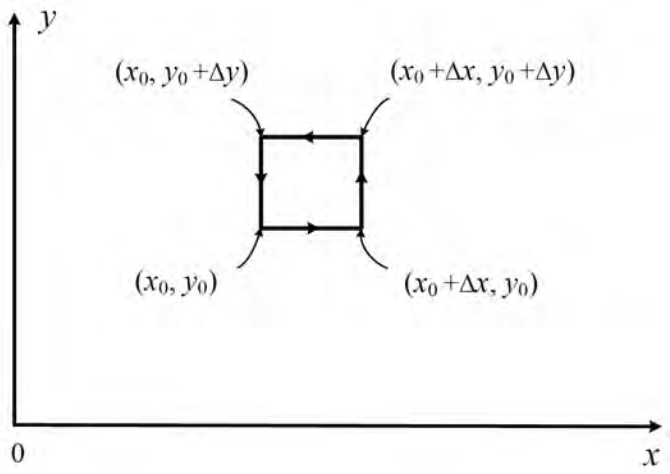


Figure 2.6: We approximate the integration over a small rect using this figure. There are four edges to this small rect.

As of this point, $\nabla \times \mathbf{E}$ is not defined explicitly. Hence, we next need to prove the details of definition (2.2.2) using Figure 2.6. Performing the integral over the small rect, one gets

$$\begin{aligned}\oint_{\Delta C} \mathbf{E} \cdot d\mathbf{l} &= E_x(x_0, y_0, z_0)\Delta x + E_y(x_0 + \Delta x, y_0, z_0)\Delta y \\ &\quad - E_x(x_0, y_0 + \Delta y, z_0)\Delta x - E_y(x_0, y_0, z_0)\Delta y \\ &= \Delta x \Delta y \left(\frac{E_x(x_0, y_0, z_0)}{\Delta y} - \frac{E_x(x_0, y_0 + \Delta y, z_0)}{\Delta y} \right. \\ &\quad \left. - \frac{E_y(x_0, y_0, z_0)}{\Delta x} + \frac{E_y(x_0 + \Delta x, y_0, z_0)}{\Delta x} \right)\end{aligned} \tag{2.2.6}$$

We have picked the normal to the incremental surface ΔS to be $\hat{z}$ in the above example, and hence, the above gives rise to the identity that

$$\lim_{\Delta S \to 0} \frac{\oint_{\Delta S} \mathbf{E} \cdot d\mathbf{l}}{\Delta S} = \frac{\partial}{\partial x} E_y - \frac{\partial}{\partial y} E_x = \hat{z} \cdot \nabla \times \mathbf{E} \tag{2.2.7}$$

Picking different $\mathbf{\Delta S}$ with different orientations and normals $\hat{n}$ where $\hat{n} = \hat{x}$ or $\hat{n} = \hat{y}$, one gets

$$\frac{\partial}{\partial y} E_z - \frac{\partial}{\partial z} E_y = \hat{x} \cdot \nabla \times \mathbf{E} \tag{2.2.8}$$

$$\frac{\partial}{\partial z} E_x - \frac{\partial}{\partial x} E_z = \hat{y} \cdot \nabla \times \mathbf{E} \tag{2.2.9}$$

The above gives the x, y, and z components of $\nabla \times \mathbf{E}$. It is to be noted that $\nabla \times \mathbf{E}$ is a vector. In other words, one gets

$$\begin{aligned} \nabla \times \mathbf{E} = \hat{x} \left(\frac{\partial}{\partial y} E_z - \frac{\partial}{\partial z} E_y \right) + \hat{y} \left(\frac{\partial}{\partial z} E_x - \frac{\partial}{\partial x} E_z \right) \\ + \hat{z} \left(\frac{\partial}{\partial x} E_y - \frac{\partial}{\partial y} E_x \right) \end{aligned} \tag{2.2.10}$$

where

$$\nabla = \hat{x} \frac{\partial}{\partial x} + \hat{y} \frac{\partial}{\partial y} + \hat{z} \frac{\partial}{\partial z} \tag{2.2.11}$$

The above gives an explicit formula for $\nabla \times \mathbf{E}$ in Cartesian coordinates.

2.2.1 Physical Meaning of Curl Operator

The curl operator $\nabla \times$ is a measure of the rotation, the torque, or the circulation of a field at a point in space.[7] On the other hand, $\oint_{\Delta C} \mathbf{E} \cdot d\mathbf{l}$ is a measure of the circulation of the field $\mathbf{E}$ around the loop formed by C. To see if a field has a non-zero curl, one can imagine a paddle wheel placed in such a field. If the field is uniform, the paddle wheel will not rotate implying that the curl of a uniform field is zero. However, if the field is varying in the z direction, then the paddle wheel will rotate, implying that a non-uniform field has non-zero curl.

[7] In many old textbook, the notation "rot" is still used for the curl or $\nabla \times$ operator.

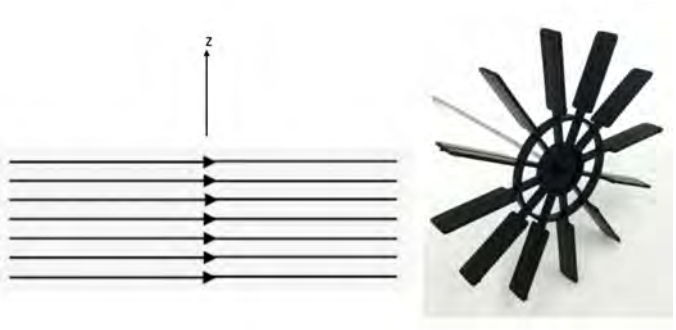


Figure 2.7: A paddle wheel can be used to test if a field has zero or non-zero curl as explained in the text.

Again, the curl operator has its complicated representations in other coordinate systems like cylindrical or spherical coordinates, a subject that will not be discussed in detail here [36].

It is to be noted that our proof of the Stokes's theorem is for a flat open surface S, and not for a general curved open surface. Since all curved surfaces can be tessellated into a union of flat triangular surfaces according to the tiling theory of simplices,[8] the generalization of the above proof to curved surface is straightforward. An example of such a triangulation of a curved surface into a union of flat triangular surfaces is shown in Figure 2.8.

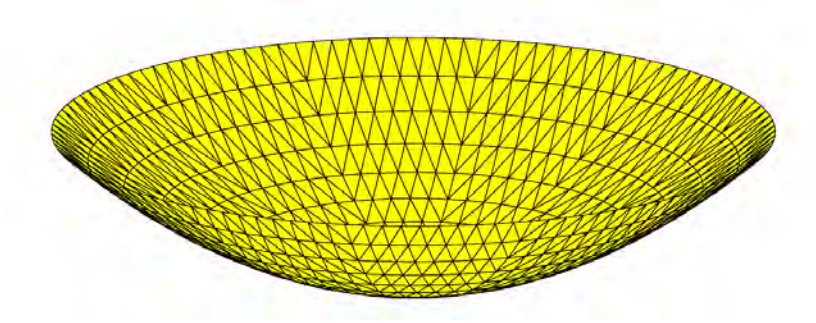


Figure 2.8: An arbitrary curved surface can be triangulated with flat triangular patches, called simplices. The triangulation can be made arbitrarily accurate by making the patches arbitrarily small.

[8]It says that any curve in 1D can be approximated by union of line segments, a 2D surface can be approximated by union of triangles, while a 3D volume can be approximated by union of tetrahedrons. Line segments, triangles, and tetrahedrons are simplices in 1D, 2D, and 3D.

2.2.2 Faraday's Law in Differential Operator Form

Faraday's law in integral form is given by[9]

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\frac{d}{dt} \iint_S \mathbf{B} \cdot d\mathbf{S} \tag{2.2.12}$$

Assuming that the surface S is not time varying, one can take the time derivative into the integrand and rewrite the above as

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\iint_S \frac{\partial}{\partial t}\mathbf{B} \cdot d\mathbf{S} \tag{2.2.13}$$

In the above, d/dt becomes $\partial/\partial t$ inside the integrand since $\mathbf{B} = \mathbf{B}(\mathbf{r}, t)$ is a multivariable function. One can replace the left-hand side with the use of Stokes' theorem to arrive at

$$\iint_S \nabla \times \mathbf{E} \cdot d\mathbf{S} = -\iint_S \frac{\partial}{\partial t}\mathbf{B} \cdot d\mathbf{S} \tag{2.2.14}$$

In the above, $d\mathbf{S}$ is an arbitraty elemental surface which can be made very small. Then the integral can be removed, and one has

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = -\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r}, t) \tag{2.2.15}$$

The above is Faraday's law in differential operator form.

In the static limit, $\frac{\partial \mathbf{B}}{\partial t} = 0$, giving

$$\nabla \times \mathbf{E} = 0 \tag{2.2.16}$$

2.3 Maxwell's Equations in Differential Operator Form

With the use of Gauss' divergence theorem and Stokes' theorem, Maxwell's equations can now be written more elegantly in differential operator forms. They are:

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = -\frac{\partial \mathbf{B}}{\partial t}(\mathbf{r}, t) \tag{2.3.1}$$

$$\nabla \times \mathbf{H}(\mathbf{r}, t) = \frac{\partial \mathbf{D}}{\partial t}(\mathbf{r}, t) + \mathbf{J}(\mathbf{r}, t) \tag{2.3.2}$$

$$\nabla \cdot \mathbf{D}(\mathbf{r}, t) = \varrho(\mathbf{r}, t) \tag{2.3.3}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r}, t) = 0 \tag{2.3.4}$$

These equations are point-wise relations as they relate the left-hand side and right-hand side field values at a given point in space. Moreover, they are not independent of each other. For instance,

[9] Faraday's law is experimentally motivated. Michael Faraday (1791-1867) was an extraordinary experimentalist who documented this law with meticulous care. It was only decades later that a mathematical description of this law was arrived at.

one can take the divergence of the first equation (2.3.1), making use of the vector identity that $\nabla \cdot (\nabla \times \mathbf{E}) = 0$, one gets

$$-\frac{\partial \nabla \cdot \mathbf{B}}{\partial t} = 0 \rightarrow \nabla \cdot \mathbf{B} = \text{constant} \tag{2.3.5}$$

This constant corresponds to magnetic charges, and since they have not been experimentally observed, one can set the constant to zero. Thus the fourth of Maxwell's equations, (2.3.4), follows from the first (2.3.1).

Similarly, by taking the divergence of the second equation (2.3.2), and making use of the current continuity equation that

$$\nabla \cdot \mathbf{J} + \frac{\partial \varrho}{\partial t} = 0 \tag{2.3.6}$$

one can obtain the second last equation (2.3.3). Notice that in (2.3.3), the charge density ϱ can be time-varying, whereas in the previous lecture, we have "derived" this equation from Coulomb's law using electrostatic theory.

The above logic follows if $\partial/\partial t \neq 0$, and is not valid for static case. In other words, for statics, the third and the fourth equations are not derivable from the first two. Hence all four Maxwell's equations are needed for static problems. For electrodynamic problems, only solving the first two suffices.

Something is amiss in the above. If $\mathbf{J}$ is known, then solving the first two equations implies solving for four vector unknowns, $\mathbf{E}, \mathbf{H}, \mathbf{B}, \mathbf{D}$, which has 12 scalar unknowns. But there are only two vector equations or 6 scalar equations in the first two equations. Thus, one needs more equations. These are provided by the constitutive relations that we shall discuss next.

2.4 Historical Notes

There are several interesting historical notes about Maxwell.

- It is to be noted that when James Clerk Maxwell first wrote his equations down, it was in many equations and very difficult to digest [17, 40, 41]. It was an eccentric English genius Oliver Heaviside, an electrical engineer by training, together with the Maxwellians [42], who distilled those equations into their present form found in textbooks. Putatively, most cannot read Maxwell's treatise [40] beyond the first 50 pages [18].
- Maxwell wrote many poems in his short lifespan (1831-1879) and they can be found at [43].
- Also, the ancestor of James Clerk Maxwell married from the Clerk family into the Maxwell family. One of the conditions of marriage was that all the descendants of the Clerk family should adopt the family name Clerk Maxwell. That was why Maxwell was addressed as Professor Clerk Maxwell in his papers.

Exercises for Lecture 2

Problem 2-1:

(i) By going through proper flux counting, show that (2.1.9) is valid.

(ii) By going through the math carefully, starting from (2.2.6), show that (2.2.10) is correct.

(iii) Explain why Stokes' theorem can be generalized to curved surfaces.

(iv) In Section 2.3 of Chapter 2, show that for the four Maxwell's equations, equations (2.3.3) and (2.3.4) are derivable from the first two Maxwell's equations.

(v) Explain why this derivation is not valid for static electromagnetic fields.

(vi) By converting the current continuity equation into integral form, explain why it is the same as charge conservation.

Chapter 3

Constitutive Relations, Wave Equation, and Static Green's Function

Constitutive relations are important for defining the electromagnetic material properties of the media involved. Also, wave phenomenon is a major triumph of Maxwell's equations. Hence, we will study the derivation of this important phenomenon here. Moreover, to make matter simple, we will reduce the problem to electrostatics to simplify the math to introduce the concept of the Green's function. This is an important concept in electromagnetics from statics to dynamics.

As mentioned previously, for time-varying problems, only the first two of the four Maxwell's equations suffice because the latter two are derivable from the first two. In other words, the latter two of Maxwell's equations are redundant, and that electromagnetics solutions are embedded in the first two. But the first two equations have four unknowns $\mathbf{E}$, $\mathbf{H}$, $\mathbf{D}$, and $\mathbf{B}$. Hence, two more equations are needed to solve for these unknowns. These extra equations come from the constitutive relations.

3.1 Simple Constitutive Relations

The constitution relation between electric flux density $\mathbf{D}$ and the electric field $\mathbf{E}$ in free space (or vacuum) is

$$\mathbf{D} = \varepsilon_0 \mathbf{E} \tag{3.1.1}$$

where $\mathbf{D}$ has the unit of coulomb per m^2, $\mathbf{E}$ has the unit or volt per m, and ε has the unit of farad per m. The $\mathbf{E}$ can be thought of as an applied field, giving rise to flux density $\mathbf{D}$ or flux flow.

It is to be noted that $\frac{\partial \mathbf{D}}{\partial t}$ has the physical meaning of displacement current like those flowing through a capacitor. When a dielectric medium is present between two parallel plate of the capacitor, this displacement current is enhanced. Thus, when material medium is present, one

has to add the contribution to $\mathbf{D}$ by the polarization density $\mathbf{P}$ in the medium, which is a dipole density.[1] Then [33, 44, 34]

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} \tag{3.1.2}$$

The second term $\mathbf{P}$ above is due to material property, and the contribution to the electric flux due to the polarization density of the medium. It is due to the little dipole contribution because of the polar nature of the atoms or molecules that make up a medium.

By the same token, the first term $\varepsilon_0 \mathbf{E}$ can be thought of as the polarization density contribution of vacuum. Vacuum, though represents nothingness, has electrons and positrons, or electron-positron pairs lurking in it [45]. Electron is matter, whereas positron is anti-matter. In the quiescent state, they represent nothingness, but they can be polarized by an electric field $\mathbf{E}$. That also explains why electromagnetic wave can propagate through vacuum.

Many media can be approximated by linear media. Then $\mathbf{P}$ is linearly proportional to the applied field $\mathbf{E}$, or $\mathbf{P} = \varepsilon_0 \chi_0 \mathbf{E}$, or[2]

$$\begin{aligned} \mathbf{D} &= \varepsilon_0 \mathbf{E} + \varepsilon_0 \chi_0 \mathbf{E} \\ &= \varepsilon_0 (1 + \chi_0) \mathbf{E} = \varepsilon \mathbf{E}, \quad \varepsilon = \varepsilon_0 (1 + \chi_0) = \varepsilon_0 \varepsilon_r \end{aligned} \tag{3.1.3}$$

where χ_0 is the electric susceptibility . In other words, for linear material media, one can replace the vacuum permittivity ε_0 with an effective permittivity $\varepsilon = \varepsilon_0 \varepsilon_r$ where ε_r is the relative permittivity. Thus, $\mathbf{D}$ is linearly proportional to $\mathbf{E}$. In free space,[3] the permittivity ε is (experimentally determined)

$$\varepsilon = \varepsilon_0 = 8.854 \times 10^{-12} \approx \frac{10^{-8}}{36\pi} \text{ F/m(Farad/m)} \tag{3.1.4}$$

The constitutive relation between magnetic flux $\mathbf{B}$ and magnetic field $\mathbf{H}$ is given as[4]

$$\mathbf{B} = \mu \mathbf{H}, \qquad \mu \text{ is the permeability in H/m (Henry/m)} \tag{3.1.5}$$

In free space or vacuum,

$$\mu = \mu_0 = 4\pi \times 10^{-7} \text{ H/m} \tag{3.1.6}$$

In the above, $\mathbf{B}$ is the magnetic flux density in unit of weber per m or tesla, and $\mathbf{H}$ has the unit of A per m (ampere per meter), while μ has the unit of henry per m. As shall be explained later, this μ_0 is an assigned value giving it a precise value as shown above. In other materials, the permeability can be written as

$$\mu = \mu_0 \mu_r \tag{3.1.7}$$

The above can be derived using similar argument as that for relative permittivity, where the different permeability is due to the presence of magnetic dipole density in a material medium. In the above, μ_r is termed the relative permeability.

[1] Note that a dipole moment is given by $Q\ell$ where Q is its charge in coulomb and ℓ is its length in m. Hence, dipole density, or polarization density as dimension of coulomb/m^2, which is the same as that of electric flux $\mathbf{D}$.

[2] This is not the most general linear relation between $\mathbf{P}$ and $\mathbf{E}$, but the simplest one we can begin with.

[3] It is to be noted that we will use MKS or SI (systeme internationale) unit in this course. Another possible unit is the CGS unit used in many physics texts [46]

[4] Again, this is not the most general linear relation, but the simplest to begin with.

3.2 Emergence of Wave Phenomenon, Triumph of Maxwell's Equations

One of the major triumphs of Maxwell's equations is the prediction of the wave phenomenon. This was experimentally verified by Heinrich Hertz in 1888 [19], some 23 years after the completion of Maxwell's theory in 1865 [17]. To see this, we consider the first two Maxwell's equations for time-varying fields in vacuum or a source-free medium.[5] They are

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \tag{3.2.1}$$

$$\nabla \times \mathbf{H} = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \tag{3.2.2}$$

Taking the curl of (3.2.1), we have

$$\nabla \times \nabla \times \mathbf{E} = -\mu_0 \frac{\partial}{\partial t} \nabla \times \mathbf{H} \tag{3.2.3}$$

It is understood that in the above, the double curl operator implies $\nabla \times (\nabla \times \mathbf{E})$. Substituting (3.2.2) into (3.2.3), we have

$$\nabla \times \nabla \times \mathbf{E} = -\mu_0 \varepsilon_0 \frac{\partial^2}{\partial t^2} \mathbf{E} \tag{3.2.4}$$

The above is the vector wave equation. In the above, the left-hand side can be simplified by using the identity that $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b})$,[6] but be mindful that the operator ∇ has to operate on a function to its right. Therefore, we arrive at the identity that

$$\nabla \times \nabla \times \mathbf{E} = \nabla \nabla \cdot \mathbf{E} - \nabla^2 \mathbf{E} \tag{3.2.5}$$

Since $\nabla \cdot \mathbf{E} = 0$ in a source-free medium, we have (3.2.4) becoming

$$\nabla^2 \mathbf{E} - \mu_0 \varepsilon_0 \frac{\partial^2}{\partial t^2} \mathbf{E} = 0 \tag{3.2.6}$$

where

$$\nabla^2 = \nabla \cdot \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

The above is known as the Laplacian operator. Here, (3.2.6) is the wave equation in three space dimensions [34, 47].

To see the simplest form of wave emerging in the above, we can let $\mathbf{E} = \hat{x} E_x(z,t)$ so that $\nabla \cdot \mathbf{E} = 0$ which is required in the source-free medium. Then (3.2.6) becomes

$$\frac{\partial^2}{\partial z^2} E_x(z,t) - \mu_0 \varepsilon_0 \frac{\partial^2}{\partial t^2} E_x(z,t) = 0 \tag{3.2.7}$$

[5] Since the third and the fourth Maxwell's equations are derivable from the first two when $\partial/\partial_t \neq 0$.

[6] For mnemonics, this formula is also known as the "back-of-the-cab" or "bac-cab" formula. It can be proved easily by expanding the vectors in their cartesian components and going through the algebraic manipulation.

Eq. (3.2.7) is known mathematically as the wave equation in one dimensional space. It can also be written as

$$\frac{\partial^2}{\partial z^2} f(z,t) - \frac{1}{c_0^2}\frac{\partial^2}{\partial t^2} f(z,t) = 0 \tag{3.2.8}$$

where $c_0^2 = (\mu_0 \varepsilon_0)^{-1}$. Eq. (3.2.8) can also be factorized as

$$\left(\frac{\partial}{\partial z} - \frac{1}{c_0}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial z} + \frac{1}{c_0}\frac{\partial}{\partial t}\right) f(z,t) = 0 \tag{3.2.9}$$

or

$$\left(\frac{\partial}{\partial z} + \frac{1}{c_0}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial z} - \frac{1}{c_0}\frac{\partial}{\partial t}\right) f(z,t) = 0 \tag{3.2.10}$$

The above can be verified easily by direct expansion, and using the fact that

$$\frac{\partial}{\partial t}\frac{\partial}{\partial z} = \frac{\partial}{\partial z}\frac{\partial}{\partial t} \tag{3.2.11}$$

The above implies that we have either

$$\left(\frac{\partial}{\partial z} + \frac{1}{c_0}\frac{\partial}{\partial t}\right) f_+(z,t) = 0 \tag{3.2.12}$$

or

$$\left(\frac{\partial}{\partial z} - \frac{1}{c_0}\frac{\partial}{\partial t}\right) f_-(z,t) = 0 \tag{3.2.13}$$

Equation (3.2.12) and (3.2.13) are known as the one-way wave equations or the advective equations [48]. From the above factorization, it is seen that the solutions to these one-way wave equations are also the solutions of the original wave equation given by (3.2.8). Their general solutions are then of the form

$$f_+(z,t) = F_+(z - c_0 t) \tag{3.2.14}$$

$$f_-(z,t) = F_-(z + c_0 t) \tag{3.2.15}$$

We can verify the above by back substitution into (3.2.12) and (3.2.13). Eq. (3.2.14) constitutes a right-traveling wave function of any shape while (3.2.15) constitutes a left-traveling wave function of any shape. Since Eqs. (3.2.14) and (3.2.15) are also solutions to (3.2.8), we can write the general solution to the wave equation as

$$f(z,t) = F_+(z - c_0 t) + F_-(z + c_0 t) \tag{3.2.16}$$

This is a wonderful result since F_+ and F_- are arbitrary functions[7] of any shape (see Figure 3.1); they can be used to encode information for communication as has happened in wireless communication which has transformed the modern world!

[7] Single value functions where their derivative exists.

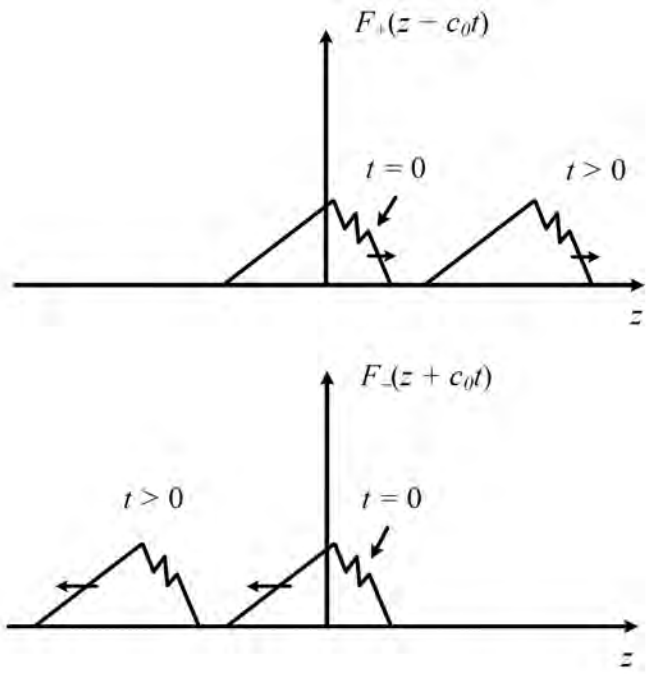


Figure 3.1: Solutions of the wave equation can be a single-valued function of any shape. In the above, F_+ travels in the positive z direction, while F_- travels in the negative z direction as t increases.

Furthermore, one can calculate the velocity of this wave to be

$$c_0 = 299{,}792{,}458\text{m/s} \simeq 3 \times 10^8\text{m/s} \tag{3.2.17}$$

where $c_0 = \sqrt{1/\mu_0\varepsilon_0}$. It is to be noted that the value of ε and μ can be changed by working in different units, but the velocity of light cannot be changed [49][p. 781].

Maxwell's equations (3.2.1) and (3.2.2) imply that $\mathbf{E}$ and $\mathbf{H}$ are linearly proportional to each other. One can define a new magnetic field variable $\mathbf{H} = \alpha\mathbf{H}'$ and see that the constants in these two equations will change. The resulting equations are

$$\nabla \times \mathbf{E} = -\mu_0\alpha\frac{\partial \mathbf{H}'}{\partial t} \tag{3.2.18}$$

$$\nabla \times \mathbf{H}' = \frac{\varepsilon_0}{\alpha}\frac{\partial \mathbf{E}}{\partial t} \tag{3.2.19}$$

One can think of the above Maxwell's equations as having a new $\mu_0' = \mu_0\alpha$, and $\varepsilon_0' = \frac{\varepsilon_0}{\alpha}$. However, upon eliminating the magnetic field from these two equations, the resulting equation is still (3.2.4) with a universal constant proportional to $\mu_0\varepsilon_0 = 1/c_0^2$ which is the velocity of light. Thus, there is only one independent constant to be determined in the wave equation which is c_0. The value

of μ_0 is defined neatly to be $4\pi \times 10^{-7}$ henry m^{-1}, while the value of ε_0 has been measured to be about 8.854×10^{-12} farad m^{-1}. Now it has been decided that the velocity of light is used as a standard and is defined to be the integer given in (3.2.17). A meter is now defined to be the distance traveled by light in 1/(299792458) second. Hence, the more accurate that unit of time or second that can be calibrated, the more accurate can we calibrate the unit of length or meter. Thus, the design of an accurate clock like an atomic clock is an important research area.

The value of ε_0 was measured in the laboratory quite early. Then it was realized that electromagnetic wave propagates at a tremendous velocity which is the velocity of light.[8] This was also the defining moment which revealed that the field of electricity and magnetism, and the field of optics were both described by Maxwell's equations or electromagnetic theory!

3.3 Static Electromagnetics–Revisted

We have seen static electromagnetics previously in integral form. Now we look at them in differential operator form. When the fields and sources are not time varying, namely that $\partial/\partial t = 0$, we arrive at the static Maxwell's equations for electrostatics and magnetostatics, namely [34, 33, 50]

$$\nabla \times \mathbf{E} = 0 \tag{3.3.1}$$

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{3.3.2}$$

$$\nabla \cdot \mathbf{D} = \varrho \tag{3.3.3}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{3.3.4}$$

Notice the the electrostatic field system is decoupled from the magnetostatic field system. However, in a resistive system where

$$\mathbf{J} = \sigma \mathbf{E} \tag{3.3.5}$$

the two systems are coupled again. This is known as resistive coupling between them. But if $\sigma \to \infty$, in the case of a perfect conductor, or superconductor, then for a finite $\mathbf{J}$, $\mathbf{E}$ has to be zero. The two systems are decoupled again.

Also, one can arrive at the equations above by letting $\mu_0 \to 0$ and $\epsilon_0 \to 0$. In this case, the velocity of light becomes infinite, or retardation (or delay) effect is negligible. In other words, it takes no time(instantaneous) for signal propagation through the system in the static approximation.

Finally, it is important to note that in statics, the latter two Maxwell's equations are not derivable from the first two. Hence, all four equations have to be considered when one seeks solutions in the static regime or the long-wavelength regime.

3.3.1 Electrostatics

Faraday's law in the static limit is

$$\nabla \times \mathbf{E} = 0 \tag{3.3.6}$$

[8] The velocity of light was known long ago in astronomy by Roemer (1676) [20].

One way to satisfy the above is to let $\mathbf{E} = -\nabla\Phi$ because of the identity $\nabla \times \nabla = 0$.[9] Alternatively, one can assume that $\mathbf{E}$ is a constant to satisfy (3.3.6). But we usually are interested in solutions that vanish at infinity, and hence, the latter is not a viable solution. Therefore, we let

$$\mathbf{E} = -\nabla\Phi \tag{3.3.7}$$

where Φ is the scalar potential. Furthermore, (3.3.1) has no unique solution. To obtain a unique solution for Φ, (3.3.1) has to be solved in tandem with (3.3.3) which we shall discuss next.

3.3.2 Electrostatics and KVL

Kirchhoff voltage law (KVL) is related to electrostatics. By applying Stokes' theorem to (3.3.6), one obtains that

$$\oint_C \mathbf{E} \cdot \mathbf{dl} = 0 \tag{3.3.8}$$

The above integration around a loop can be written as a sum of integral over line segments. Since $\mathbf{E} = -\nabla\Phi$, the integral over a line segment can be performed in closed form yielding that

$$\int_a^b \mathbf{E} \cdot \mathbf{dl} = -\Phi_b - \Phi_a \tag{3.3.9}$$

Hence, the (3.3.8) implies that the potential-drop around a loop is zero, a statement of KVL.

3.3.3 Poisson's Equation

As a consequence of the above discussion,

$$\nabla \cdot \mathbf{D} = \varrho \Rightarrow \nabla \cdot \varepsilon\mathbf{E} = \varrho \Rightarrow -\nabla \cdot \varepsilon\nabla\Phi = \varrho \tag{3.3.10}$$

In the last equation above, if ε is a constant of space, or independent of $\mathbf{r}$, then ε and $\nabla\cdot$ commute, or ε can be moved to the left of $\nabla\cdot$, and one arrives at the simple Poisson's equation, which is a partial differential equation

$$\nabla^2\Phi = -\frac{\varrho}{\varepsilon} \tag{3.3.11}$$

Here, the Laplacian operator

$$\nabla^2 = \nabla \cdot \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

Solving a partial differential equation is quite involved, but we can find a simple solution to the above partial differential equation in the following simple case. For a point source, we know from Coulomb's law that

$$\mathbf{E} = \frac{q}{4\pi\varepsilon r^2}\hat{r} = -\nabla\Phi \tag{3.3.12}$$

[9] One an easily go through the algebra in cartesian coordinates to convince oneself of this.

From the above, we deduce that[10]

$$\Phi = \frac{q}{4\pi\varepsilon r} \tag{3.3.13}$$

Therefore, we know the solution to Poisson's equation (3.3.11) is given by the above when the source ϱ represents a point source in the above.

Since this is a *linear equation*, we can use the principle of linear superposition to find the solution when the charge density $\varrho(\mathbf{r})$ is arbitrary. To this end, a point source located at $\mathbf{r}'$ is described by a charge density as

$$\varrho(\mathbf{r}) = q\delta(\mathbf{r} - \mathbf{r}') \tag{3.3.14}$$

where $\delta(\mathbf{r} - \mathbf{r}')$ is a short-hand notation for $\delta(x - x')\delta(y - y')\delta(z - z')$. Therefore, from (3.3.11), the corresponding partial differential equation for a point source is

$$\nabla^2\Phi(\mathbf{r}) = -\frac{q\delta(\mathbf{r} - \mathbf{r}')}{\varepsilon} \tag{3.3.15}$$

The solution to the above equation, from Coulomb's law in accordance to (3.3.13), has to be

$$\Phi(\mathbf{r}) = \frac{q}{4\pi\varepsilon|\mathbf{r} - \mathbf{r}'|} \tag{3.3.16}$$

whereas (3.3.13) is for a point source at the origin, but (3.3.16) is for a point source located at $\mathbf{r} = \mathbf{r}'$. The above is a coordinate independent form of the solution. Here, we adopt the notation that $\mathbf{r} = \hat{x}x + \hat{y}y + \hat{z}z$ and $\mathbf{r}' = \hat{x}x' + \hat{y}y' + \hat{z}z'$, and $|\mathbf{r} - \mathbf{r}'| = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}$.

3.3.4 Static Green's Function

The response to a linear time-invariant system driven by an impulse function is known as the impulse response in the time domain. In the space domain, an analogous solution when the source is a point source is called the Green's function.

Let us define a partial differential equation given by

$$\nabla^2 g(\mathbf{r} - \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}') \tag{3.3.17}$$

The above is similar to Poisson's equation with a point source on the right-hand side as in (3.3.15). Such a solution, which is a response to a point source, is called the Green's function.[11] By comparing equations (3.3.15) and (3.3.17), then making use of (3.3.16), we deduced that the static Green's function is

$$g(\mathbf{r} - \mathbf{r}') = \frac{1}{4\pi|\mathbf{r} - \mathbf{r}'|} \tag{3.3.18}$$

An arbitrary source, using the sifting property of a delta function, it can be expressed as

$$\varrho(\mathbf{r}) = \iiint_V dV' \varrho(\mathbf{r}')\delta(\mathbf{r} - \mathbf{r}') \tag{3.3.19}$$

[10] One can always take the gradient or ∇ of Φ to verify this. Mind you, this is best done in spherical coordinates.

[11] George Green (1793-1841), the son of a Nottingham miller, was self-taught, but his work has a profound impact in our world.

where V is the volume over which the charge density $\varrho(\mathbf{r}')$ is nonzero. It is also called the support of the charge density. The above is just the statement that an arbitrary charge distribution $\varrho(\mathbf{r})$ can be expressed as a linear superposition of point sources $\delta(\mathbf{r}-\mathbf{r}')$. Using the above in (3.3.11), we have

$$\nabla^2\Phi(\mathbf{r}) = -\frac{1}{\varepsilon}\iiint_V dV'\varrho(r')\delta(\mathbf{r}-\mathbf{r}') \tag{3.3.20}$$

Hence, we can let

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon}\iiint_V dV' g(\mathbf{r}-\mathbf{r}')\varrho(\mathbf{r}') \tag{3.3.21}$$

By substituting the above into the left-hand side of (3.3.20), exchanging order of integration and differentiation, and then making use of equation (3.3.17), it can be shown that (3.3.21) indeed satisfies (3.3.11). The above is just a 3D convolutional integral. Hence, the potential $\Phi(\mathbf{r})$ due to an arbitrary source distribution $\varrho(\mathbf{r})$ can be found by using 3D convolution, namely,

$$\Phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon}\iiint_V \frac{\varrho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}dV' \tag{3.3.22}$$

In a nutshell, the solution of Poisson's equation when it is driven by an arbitrary source ϱ, is the 3D convolution of the source $\varrho(\mathbf{r})$ with the static Green's function $g(\mathbf{r})$, whichi is a point source response. In the above, $dV' = dx'dy'dz'$ which is a 3D integration. It is also variously written as $d\mathbf{r}'$ or $d^3\mathbf{r}'$.

3.3.5 Laplace's Equation

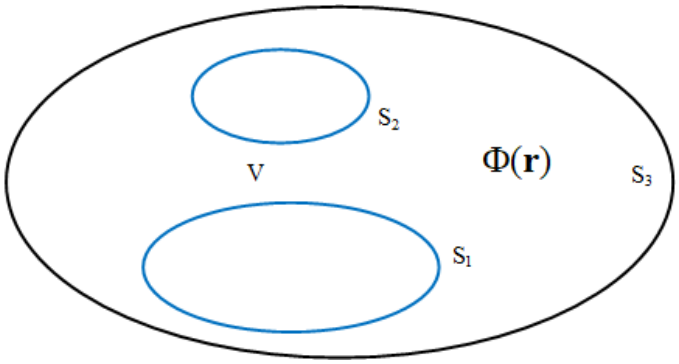


Figure 3.2: A boundary value problem (BVP) for a partial differential equation is usually solved in a region V bounded by a surface S. In this case, the volume V is bounded by three surfaces S_1, S_2, and S_3. The field $\Phi(\mathbf{r})$ needs to satisfy the partial differential equation inside the volume V. Boundary conditions are specified for the field $\Phi(\mathbf{r})$ to be satisfied on disconnected surfaces S_1, S_2, and S_3.

If $\varrho = 0$, or if we are in a source-free region, then for electrostatics,

$$\nabla^2 \Phi = 0 \tag{3.3.23}$$

which is the Laplace's equation with $\Phi = 0$ in the trivial case. For the non-trivial solution, Laplace's equation is usually solved as a boundary value problem. In such a problem, the potential $\Phi(\mathbf{r})$ needs to be found inside a volume V bounded by a surface S. The value of $\Phi(\mathbf{r})$ is stipulated on the boundary of a region with a certain boundary condition, and then the solution is sought in the volume V so as to match the boundary condition on the surface S. Examples of such boundary value problems are given at the end of the lecture.

Exercises for Lecture 3

Problem 3-1:

(i) Show that (3.2.14) and (3.2.15) are solutions to (3.2.12) and (3.2.13), respectively.

(ii) For static electromagnetics, explain why when a resistive medium exists, the electrostatic system is not decoupled from the magnetostatic system.

(iii) Explain why (3.3.16) is the solution to (3.3.15).

Chapter 4

Magnetostatics, Boundary Conditions, and Jump Conditions

In the previous lecture, Maxwell's equations become greatly simplified in the static limit. We have looked at how the electrostatic problems are solved. We now look at the magnetostatic case. In addition, we will study boundary conditions and jump conditions at an interface, and how they can be derived from Maxwell's equations. Maxwell's equations can be first solved in different physical regions. Then the solutions are pieced (or sewn) together by imposing boundary conditions at the boundaries or interfaces of the regions. Such problems are called boundary-value problems (BVPs).

4.1 Magnetostatics

From Maxwell's equations, we can deduce that the magnetostatic equations for the magnetic field and flux when $\partial/\partial t = 0$, are [50, 34, 33]

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{4.1.1}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{4.1.2}$$

Both the equations above do not have unique solutions. Hence, to solve the above, we have to invoke the constitutive relation that $\mathbf{B} = \mu\mathbf{H}$. These two equations are greatly simplified, and hence, are easier to solve compared to the time-varying case. One way to satisfy the second equation is to let

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{4.1.3}$$

In the above, (4.1.2) does not have a unique solution. By letting $\mathbf{B}$ as in (4.1.3), because of the vector identity

$$\nabla \cdot (\nabla \times \mathbf{A}) = 0 \tag{4.1.4}$$

(4.1.2) is always satisfied. (4.1.4) is zero for the same algebraic reason that $\mathbf{a}\cdot(\mathbf{a}\times\mathbf{b})=0$. Here, $\mathbf{A}$ is also known as the magnetic vector potential for magnetic field analogous to that Φ is the scalar potential for electric field. In this manner, Gauss's law in (4.1.2) is automatically satisfied.

From (4.1.1), we have

$$\nabla\times\left(\frac{\mathbf{B}}{\mu}\right)=\mathbf{J} \tag{4.1.5}$$

Then using (4.1.3) into the above, we get

$$\nabla\times\left(\frac{1}{\mu}\nabla\times\mathbf{A}\right)=\mathbf{J} \tag{4.1.6}$$

In a homogeneous medium,[1] μ or $1/\mu$ is a constant and it commutes with the differential ∇ operator or that it can be taken outside the differential operator. As such, one arrives at

$$\nabla\times(\nabla\times\mathbf{A})=\mu\mathbf{J} \tag{4.1.7}$$

We use the vector identity that (see BAC-CAB formula in the previous lecture)

$$\begin{aligned}\nabla\times(\nabla\times\mathbf{A})&=\nabla(\nabla\cdot\mathbf{A})-(\nabla\cdot\nabla)\mathbf{A}\\&=\nabla(\nabla\cdot\mathbf{A})-\nabla^2\mathbf{A}\end{aligned} \tag{4.1.8}$$

where ∇^2 is a shorthand notation for $\nabla\cdot\nabla$. The above is best understood in the cartesian coordinates. As a result, we arrive at [51]

$$\nabla(\nabla\cdot\mathbf{A})-\nabla^2\mathbf{A}=\mu\mathbf{J} \tag{4.1.9}$$

By imposing the Coulomb gauge that $\nabla\cdot\mathbf{A}=0$, (which will be elaborated in the next section) we arrive at the simplified equation

$$\nabla^2\mathbf{A}=-\mu\mathbf{J} \tag{4.1.10}$$

The above is also known as the vector Poisson's equation, which is a partial differential equation. In cartesian coordinates, the above can be viewed as three scalar Poisson's equations. Each of the Poisson's equation can be solved using the Green's function method previously described. Consequently, in free space

$$\mathbf{A}(\mathbf{r})=\frac{\mu}{4\pi}\iiint_V\frac{\mathbf{J}(\mathbf{r}')}{R}dV' \tag{4.1.11}$$

where

$$R=|\mathbf{r}-\mathbf{r}'| \tag{4.1.12}$$

is the distance between the source point $\mathbf{r}'$ and the observation point $\mathbf{r}$.

[1] Its prudent to warn the reader of the use of the word "homogeneous". In the math community, it usually refers to something to be set to zero. But in the electromagnetics community, it refers to something "non-heterogeneous".

4.1.1 More on Coulomb Gauge

Gauge is a very important concept in physics [49], and we will further elaborate it here. First, notice that $\mathbf{A}$ in (4.1.3) is not unique because one can always define

$$\mathbf{A}' = \mathbf{A} - \nabla\Psi \tag{4.1.13}$$

Then

$$\nabla \times \mathbf{A}' = \nabla \times (\mathbf{A} - \nabla\Psi) = \nabla \times \mathbf{A} = \mathbf{B} \tag{4.1.14}$$

where we have made use of that $\nabla \times \nabla\Psi = 0$. Hence, the $\nabla\times$ of both $\mathbf{A}$ and $\mathbf{A}'$ produce the same $\mathbf{B}$, implying that $\mathbf{A}$ is non-unique.

To find $\mathbf{A}$ uniquely, we have to define or set the divergence of $\mathbf{A}$ or provide a gauge condition or gauge fixing. One way is to set the divergence of $\mathbf{A}$ to be zero, namely that

$$\nabla \cdot \mathbf{A} = 0 \tag{4.1.15}$$

This will pin the value of $\mathbf{A}$. Then from (4.1.13)

$$\nabla \cdot \mathbf{A}' = \nabla \cdot \mathbf{A} - \nabla^2\Psi \neq \nabla \cdot \mathbf{A} \tag{4.1.16}$$

The last non-equal sign follows if $\nabla^2\Psi \neq 0$. However, if we impose the Coulomb gauge always when seeking the solutions, or we further stipulate that $\nabla \cdot \mathbf{A}' = \nabla \cdot \mathbf{A} = 0$, then $-\nabla^2\Psi = 0$. (This does not necessary imply that $\nabla\Psi = 0$, but if we impose that condition that $\Psi \to$ constant when $\mathbf{r} \to \infty$, then $\Psi =$ constant everywhere.[2]) By so doing, $\mathbf{A}$ and $\mathbf{A}'$ are equal to each other implying uniqueness, and we obtain (4.1.10) and (4.1.11).

To see this more simply, the above is akin to the idea that given a vector $\mathbf{a}$, just by stipulating that $\mathbf{b} \times \mathbf{a} = \mathbf{c}$ is not enough to determine $\mathbf{a}$. We need to stipulate as well what $\mathbf{b} \cdot \mathbf{a}$ is. Here, $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are independent vectors. Another way of saying this is that the vector $\mathbf{a}$ can be written as $\mathbf{a} = \mathbf{a}_\parallel + \mathbf{a}_\perp$ where $\mathbf{a}_\parallel$ is parallel to the vector $\mathbf{b}$, while $\mathbf{a}_\perp$ is perpendicular to $\mathbf{b}$. Here, $\mathbf{a}_\parallel$ is indeterminate now, since $\mathbf{b} \times \mathbf{a}_\parallel = 0$. But by letting $\mathbf{b} \cdot \mathbf{a} = 0$ will force $\mathbf{a}_\parallel = 0$ and define $\mathbf{a}$ uniquely.

4.1.2 Magnetostatics and KCL

The Kirchhoff current law (KCL) is intimately related to magnetostatics. If one takes the divergence of (4.1.1), one gets $\nabla \cdot \mathbf{J} = 0$. This says that the net current flowing into a point is zero, which is that statement of KCL. One can also integrate this equation about a small volume, and use Gauss' divergence theorem to relate it to a surface integral. Then KCL is obviated by such a formula.

[2] This is the same statement that a monopole-only line source is absent. All sources are finite or dipolar so that current can flow between the two poles of the source. Then it is a property of the Laplace boundary value problem that if $\Psi = 0$ on a closed surface S, then $\Psi = 0$ everywhere inside S. Earnshaw's theorem [33] is useful for proving this assertion.

4.2 Boundary Conditions—1D Poisson's Equation

To simplify the solutions of Maxwell's equations, they are usually solved in a homogeneous medium. As mentioned before, a complex problem can be divided into piecewise homogeneous regions first, and then the solution in each region sought separately. Then the total solution must satisfy boundary conditions at the interface between the piecewise homogeneous regions.

What are these boundary conditions? Boundary conditions are actually embedded in the partial differential equations (of which Maxwell's equations and Poisson's equation are) that the potential or the field satisfy. Two important concepts to keep in mind are:

- Differentiation of a function with discontinuous slope will give rise to step discontinuity.
- Differentiation of a function with step discontinuity will give rise to a Dirac delta function. This is also called the jump condition, a term often used by the mathematics community [52].

Take for example a one dimensional Poisson's equation that

$$\frac{d}{dx}\varepsilon(x)\frac{d}{dx}\Phi(x) = -\varrho(x) \tag{4.2.1}$$

where $\varepsilon(x)$ represents material property that has the functional form given in Figure 4.1. One can actually say something about $\Phi(x)$ given $\varrho(x)$ on the right-hand side. If $\varrho(x)$ has a delta function singularity, it implies that $\varepsilon(x)\frac{d}{dx}\Phi(x)$ has a step discontinuity since the derivative of a step function is a delta function. If $\varrho(x)$ is finite everywhere, then $\varepsilon(x)\frac{d}{dx}\Phi(x)$ must be continuous everywhere.

Furthermore, if $\varepsilon(x)\frac{d}{dx}\Phi(x)$ is finite everywhere, it implies that $\Phi(x)$ must be continuous everywhere.

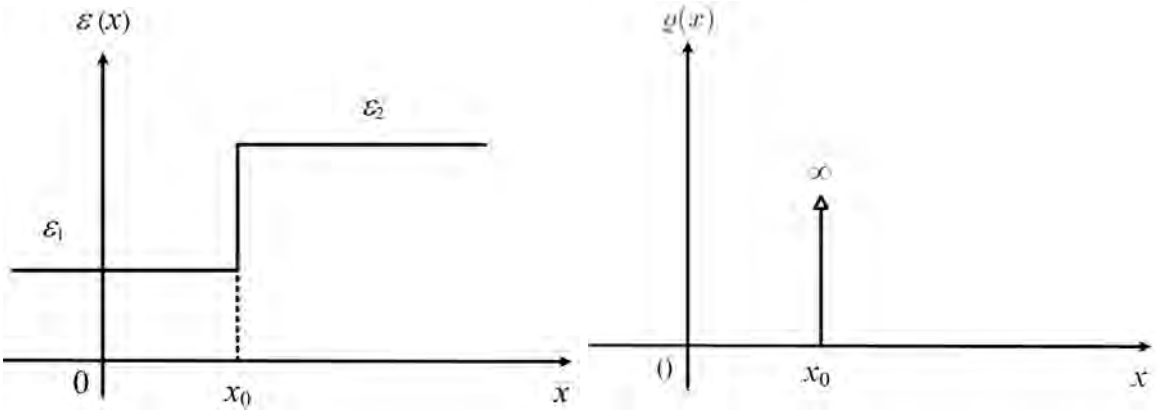


Figure 4.1: A figure showing a charge sheet $\varrho(x) = \varrho_s\delta(x - x_0)$ at the interface between two dielectric media. Because it is a surface charge sheet, the volume charge density $\varrho(x)$ is infinite at the sheet location x_0.

To see this in greater detail, we illustrate it with the following example. In the above, $\varrho(x)$ represents a singular charge distribution given by $\varrho(x) = \varrho_s\delta(x - x_0)$. In this case, the charge distribution is everywhere zero except at the location of the surface charge sheet, where the charge density is infinite: it is represented mathematically by a delta function[3] in space.

To find the boundary condition of the potential $\Phi(x)$ at x_0 , we integrate (4.2.1) over an infinitesimal width around x_0, the location of the charge sheet, namely

$$\int_{x_0-\Delta}^{x_0+\Delta} dx \left[\frac{d}{dx}\varepsilon(x)\frac{d}{dx}\Phi(x)\right] = -\int_{x_0-\Delta}^{x_0+\Delta} dx\varrho(x) = -\int_{x_0-\Delta}^{x_0+\Delta} dx\varrho_s\delta(x - x_0) \tag{4.2.2}$$

Since the integrand of the left-hand side is an exact derivative, we get

$$\varepsilon(x)\frac{d}{dx}\Phi(x)\Big|_{x_0-\Delta}^{x_0+\Delta} = -\varrho_s \tag{4.2.3}$$

whereas on the right-hand side, we pick up the contribution from the delta function. Evaluating the left-hand side at their limits, one arrives at

$$\varepsilon(x_0^+)\frac{d}{dx}\Phi(x_0^+) - \varepsilon(x_0^-)\frac{d}{dx}\Phi(x_0^-) \cong -\varrho_s, \tag{4.2.4}$$

where $x_0^{\pm} = \lim_{\Delta\to 0} x_0 \pm \Delta$. In other words, the jump discontinuity is in $\varepsilon(x)\frac{d}{dx}\Phi(x)$ and the amplitude of the jump discontinuity is proportional to the amplitude of the delta function, ϱ_s.

Since $\mathbf{E} = -\nabla\Phi$, then

$$E_x(x) = -\frac{d}{dx}\Phi(x), \tag{4.2.5}$$

The above, together with (4.2.4), implies the boundary condition that

$$\varepsilon(x_0^+)E_x(x_0^+) - \varepsilon(x_0^-)E_x(x_0^-) = \varrho_s \tag{4.2.6}$$

or

$$D_x(x_0^+) - D_x(x_0^-) = \varrho_s \tag{4.2.7}$$

where

$$D_x(x) = \varepsilon(x)E_x(x) \tag{4.2.8}$$

If $\varrho_s = 0$, then the boundary condition becomes $D_x(x_0^+) = D_x(x_0^-)$.

The lesson learned from above is that boundary condition is obtained by integrating the pertinent differential equation over an infinitesimal small segment. In this mathematical way of looking at the boundary condition, one can also eyeball the differential equation and ascertain the terms that will have the jump discontinuity and whose derivatives will yield the delta function on the right-hand side.

[3]This function has been attributed to Dirac who used in pervasively, but Cauchy was aware of such a function.

4.3 Boundary Conditions—Maxwell's Equations

As seen previously, boundary conditions for a field is embedded in the differential equation, or in general, the partial differential equation, that the field satisfies. Hence, boundary conditions can be derived from the differential operator forms of Maxwell's equations. In most textbooks, boundary conditions are obtained by integrating Maxwell's equations over a small pill box [34, 51, 33]. To derive these boundary conditions, we will take an unconventional view: namely to see what sources can induce jump conditions (or jump discontinuities) on the pertinent fields. Boundary conditions are needed at media interfaces, as well as across current or charge sheets. As shall be shown, each of the Maxwell's equations induces a boundary condition at the interface between two media or two regions separated by surface sources which are infinitely thin. Hopefully, this will give you more physical insight into the reasons for these boundary conditions.

4.3.1 Faraday's Law

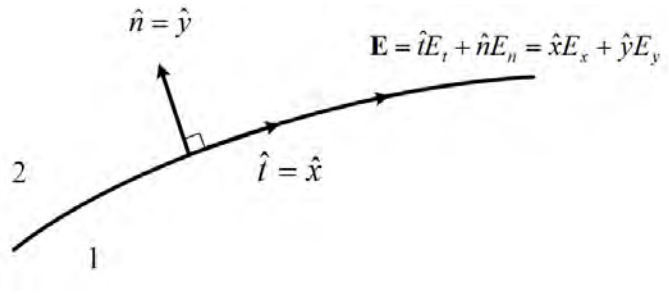


Figure 4.2: This figure is for the derivation of boundary condition induced by Faraday's law. A local coordinate system can be used to see the boundary condition more lucidly. Here, the normal $\hat{n} = \hat{y}$ and the tangential component $\hat{t} = \hat{x}$.

For this, we start with Faraday's law, which implies that

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{4.3.1}$$

The right-hand side of this equation is a derivative of a time-varying magnetic flux, which is a finite quantity. One quick answer we could ask is that if the right-hand side of the above equation is everywhere finite, could there be any jump discontinuity on the field $\mathbf{E}$ on the left hand side? The answer is a resounding no!

To see this quickly, one can project the tangential field component and normal field component to a local coordinate system. In other words, one can think of $\hat{t}$ and $\hat{n}$ as the local $\hat{x}$ and $\hat{y}$

coordinates. Then writing the curl operator in this local coordinates, one gets

$$\begin{aligned}\nabla \times \mathbf{E} &= \left(\hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y}\right) \times (\hat{x}E_x + \hat{y}E_y) \\ &= \hat{z}\frac{\partial}{\partial x}E_y - \hat{z}\frac{\partial}{\partial y}E_x \end{aligned} \tag{4.3.2}$$

In simplifying the above, we have used the distributive property of cross product, and evaluating the cross product in cartesian coordinates. The cross product gives four terms, but only two of the four terms are non-zero as shown above.

Since the right-hand side of (4.3.1) is finite, the above implies that $\frac{\partial}{\partial x}E_y$ and $\frac{\partial}{\partial y}E_x$ have to be finite. In order words, E_x is continuous in the y direction and E_y is continuous in the x direction. Since in the local coordinate system, $E_x = E_t$, then E_t is continuous across the boundary. The above implies that

$$E_{1t} = E_{2t} \tag{4.3.3}$$

or the tangential components of the electric field is continuous at the interface. To express this in a compact coordinate independent manner, we have

$$\hat{n} \times \mathbf{E}_1 = \hat{n} \times \mathbf{E}_2 \tag{4.3.4}$$

where $\hat{n}$ is the unit normal at the interface, and $\hat{n} \times \mathbf{E}$ always extracts the tangential component of a vector $\mathbf{E}$ (convince yourself).

4.3.2 Gauss's Law for Electric Flux

From Gauss's law, we have

$$\nabla \cdot \mathbf{D} = \varrho \tag{4.3.5}$$

where ϱ is the volume charge density. We would like to express this equation at an interface in terms of a local self-coordinate system.

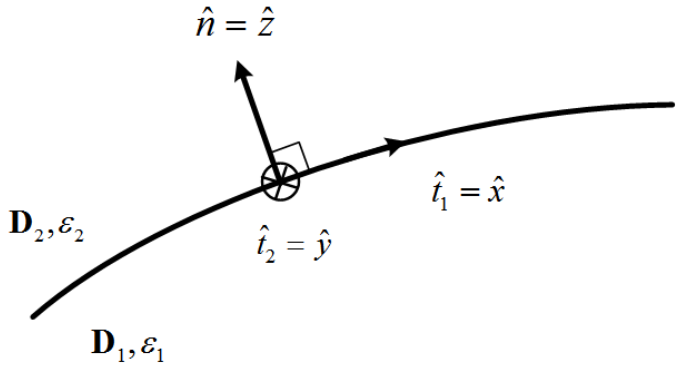


Figure 4.3: A figure showing the derivation of boundary condition for Gauss's law. Again, a local coordinate system can be introduced for simplicity.

Expressing the above in local coordinates (x, y, z as shown in Figure 4.3), then

$$\nabla \cdot \mathbf{D} = \frac{\partial}{\partial x} D_x + \frac{\partial}{\partial y} D_y + \frac{\partial}{\partial z} D_z = \varrho \tag{4.3.6}$$

The boundary condition for the electric flux can be found by *singularity matching*. If there is a surface layer charge at the interface, then the volume charge density must be infinitely large or singular; hence, it can be expressed in terms of a delta function, or $\varrho = \varrho_s \delta(z)$ in local coordinates. Here, ϱ_s is the surface charge density. By looking at the above expression in local coordinates, the only term that can possibly produce a $\delta(z)$ is from $\frac{\partial}{\partial z} D_z$. In other words, D_z must have a jump discontinuity at $z = 0$; the other terms do not. Then

$$\frac{\partial}{\partial z} D_z = \varrho_s \delta(z) \tag{4.3.7}$$

Integrating the above from $0 - \Delta$ to $0 + \Delta$, we get

$$D_z(z)\Big|_{0-\Delta}^{0+\Delta} = \varrho_s \tag{4.3.8}$$

or in the limit when $\Delta \to 0$,

$$D_z(0^+) - D_z(0^-) = \varrho_s \tag{4.3.9}$$

where $0^+ = \lim_{\Delta \to 0} 0 + \Delta$, and $0^- = \lim_{\Delta \to 0} 0 - \Delta$. Since $D_z(0^+) = D_{2n}$, $D_z(0^-) = D_{1n}$, (where D_{in} is the normal field at the i-th interface) the above becomes

$$D_{2n} - D_{1n} = \varrho_s \tag{4.3.10}$$

In other words, physically, a charge sheet ϱ_s can give rise to a jump discontinuity in the normal component of the electric flux $\mathbf{D}$. Expressed in a compact, coordinate independent form, the boundary condition is

$$\hat{n} \cdot (\mathbf{D}_2 - \mathbf{D}_1) = \varrho_s \tag{4.3.11}$$

Using the physical notion that an electric charge has electric flux $\mathbf{D}$ exuding from it, Figure 4.4 shows an intuitive sketch as to why a charge sheet gives rise to a discontinuous normal component of the electric flux $\mathbf{D}$.

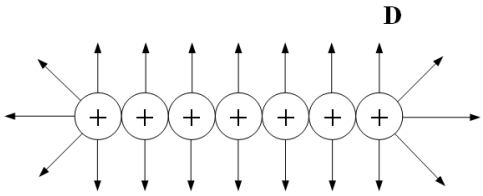


Figure 4.4: A figure intuitively showing why a sheet of electric charge gives rise to a jump discontinuiy in the normal component of the electric flux $\mathbf{D}$.

4.3.3 Ampere's Law

Ampere's law, or the generalized one, stipulates that

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{4.3.12}$$

Again if the right-hand side is everywhere finite, then $\mathbf{H}$ is a continuous field everywhere. However, if the right-hand side has a delta function singularity, due to a current sheet $\mathbf{J}$ in 3D space, and that $\frac{\partial \mathbf{D}}{\partial t}$ is regular or finite everywhere, then the only place where the singularity can be matched on the left-hand side is from the derivative of the magnetic field $\mathbf{H}$ or $\nabla \times \mathbf{H}$. In a word, $\mathbf{H}$ is not continuous. For instance, we can project the above equation onto a local coordinate system just as we did for Faraday's law.

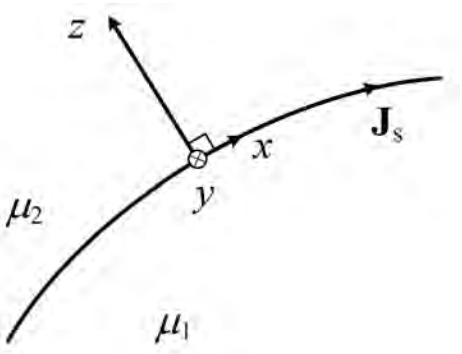


Figure 4.5: A figure showing the derivation of boundary condition for Ampere's law. A local coordinate system is used for simplicity.

To be general, we also include the presence of a current sheet at the interface. We expect the current sheet to induce a jump discontinuity in magnetic field. This is illustrated intuitively in Figure 4.6.

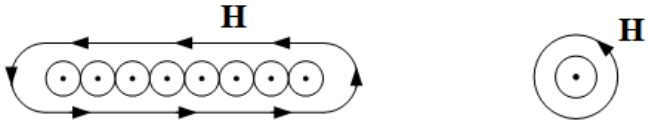


Figure 4.6: A figure intuitively showing that with the understanding of how a single line current source generates a magnetic field (right), a cluster of them forming a current sheet will generate a jump discontinuity in the tangential component of the magnetic field $\mathbf{H}$ (left).

A current sheet, or a surface current density becomes a delta function singularity when expressed as a volume current density. Thus, rewriting (4.3.12) in a local coordinate system, assuming that

$\mathbf{J} = \hat{x} J_{sx} \delta(z)$,[4] then singularity matching in local coordinates,

$$\nabla \times \mathbf{H} = \hat{x} \left(\frac{\partial}{\partial y} H_z - \frac{\partial}{\partial z} H_y \right) = \hat{x} J_{sx} \delta(z) \tag{4.3.13}$$

The displacement current term on the right-hand side is ignored since it is regular or finite, and will not induce a jump discontinuity on the field; hence, we have the form of the right-hand side of the above equation. From the above, the only term that can produce a $\delta(z)$ singularity on the left-hand side is the $-\frac{\partial}{\partial z} H_y$ term. Therefore, by singularity matching, we conclude that

$$-\frac{\partial}{\partial z} H_y = J_{sx} \delta(z) \tag{4.3.14}$$

In other words, physically, H_y has to have a jump discontinuity at the interface where the current sheet resides; or that

$$H_y(z = 0^+) - H_y(z = 0^-) = -J_{sx} \tag{4.3.15}$$

The above implies that

$$H_{2y} - H_{1y} = -J_{sx} \tag{4.3.16}$$

But H_y is just the tangential component of the $\mathbf{H}$ field. In a word, physically, the current sheet J_{sx} induces a jump discontinuity on the y component of the magnetic field. Now if we repeat the same exercise with a current with a y component, or $\mathbf{J} = \hat{y} J_{sy} \delta(z)$, at the interface, we have

$$H_{2x} - H_{1x} = J_{sy} \tag{4.3.17}$$

Now, (4.3.16) and (4.3.17) can be rewritten using a cross product as

$$\hat{z} \times (\hat{y} H_{2y} - \hat{y} H_{1y}) = \hat{x} J_{sx} \tag{4.3.18}$$

$$\hat{z} \times (\hat{x} H_{2x} - \hat{x} H_{1x}) = \hat{y} J_{sy} \tag{4.3.19}$$

The above two equations can be consolidated, written in a coordinate independent form, to give

$$\hat{n} \times (\mathbf{H}_2 - \mathbf{H}_1) = \mathbf{J}_s \tag{4.3.20}$$

where in this case here, $\hat{n} = \hat{z}$. In other words, a current sheet $\mathbf{J}_s$ can give rise to a jump discontinuity in the tangential components of the magnetic field, $\hat{n} \times \mathbf{H}$.

4.3.4 Gauss's Law for Magnetic Flux

Similarly, from Gauss's law for magnetic flux, or that

$$\nabla \cdot \mathbf{B} = 0 \tag{4.3.21}$$

[4]The form of this equation can be checked by dimensional analysis. Here, $\mathbf{J}$ has the unit of A m^{-2}, $\delta(z)$ has unit of m^{-1}, and J_{sx}, a current sheet density, has unit of A m^{-1}.

one deduces that

$$\hat{n} \cdot (\mathbf{B}_2 - \mathbf{B}_1) = 0 \tag{4.3.22}$$

or that the normal magnetic fluxes are continuous at an interface. In other words, since magnetic charges do not exist, the normal component of the magnetic flux has to be continuous.

The take-home message here is that the boundary conditions are buried in the differential operators and source singularities. If there are singular terms such as sheet sources in Maxwell's equations, then via the differential operators, the boundary conditions can be deduced. These boundary conditions at an interface are also known as jump condition if a current or a source sheet is present.

4.3.5 Locally Flat Surfaces

It is noted that the above boundary conditions are derived for locally flat surfaces. But all surfaces can be formed by union of curved surfaces. When an observation point is very close to a curved surface compared to the radius of curvature of the surface, it can be assumed to be locally flat. As we shall learn later, an electromagnetic field cannot distinguish between a sharp surface and a curved surface when the radius of curvature of the curved surface is small compared to the wavelength.

Exercises for Lecture 4

Problem 4-1:

The Earnshaw's theorem says that a minimum or maximum point cannot appear in the potential of the solution to Laplace's equation. The detail of the proof is quite long, but we can motivate this theorem in the following way:

(i) Laplace's equation can be solved by the separation of variables, namely that in 2D, its solution can be written as $\Phi(x, y) = A\cos(ax)\exp(-ay)$. Show that this is a solution to Laplace's equation, and that this function does not have a maximum or a minimum point except at the boundary.

(ii) However, if we write $\Phi(x, y) = A\cos(ax)\cos(by)$, show that this function does have a maximum or a minimum, but it is not a solution of Laplace equation. Earnshaw's theorem means that if Φ is a solution to Laplace's equation in a region V, Φ can only have maximum or minimum value at the boundary of V.

(iii) Use this to explain that if a region V is bounded by a surface S, and if Φ is constant on S, then it is the same constant everywhere in V.

(iv) Use this fact to explain how the Faraday's cage works.

(v) Explain why that Coulomb gauge can be used to guarantee a unique vector potential $\mathbf{A}$.

Problem 4-2:

(i) By back substitution, show that eq. (4.1.11) in fact satisfies (4.1.10).

(ii) Coulomb's law give the scalar potential for a monopole charge to be

$$\Phi = \frac{q}{4\pi\varepsilon r}$$

Assume that two monopoles are aligned on the z axis, and spaced infinitesimally small apart, show that by differentiating this expression with respect to z, one can get the scalar potential for a dipole to be

$$\Phi_d = \frac{\ell q\cos(\theta)}{4\pi\varepsilon r^2}$$

where ℓ is the length of the dipole.

(iii) Give the physical meaning of this mathematical procedure.

(iv) When the above Φ_d is back substituted into Poission's equation, what do you expect the charge density to be on the right-hand side of Poisson's equation?

(v) Derive the jump condition or the boundary condition induced by Ampere's law when there is a current sheet at the interface between two media.

Chapter 5

Biot-Savart law, Conductive Media Interface, Instantaneous Poynting's Theorem

Biot-Savart law, like Ampere's law was experimentally determined in around 1820 and it is discussed in a number of textbooks [49, 34, 33]. This is the cumulative work of Ampere, Oersted, Biot, and Savart. At this stage of the course, we have learnt enough mathematical tools to derive this law from Ampere's law and Gauss's law for magnetostatics. So it is appropriate at this point to show the power of mathematical logic in deriving a law inferred experimentally eons ago. In addition, we will study the boundary conditions at conductive media interfaces, and introduce the instantaneous Poynting's theorem.

5.1 Derivation of Biot-Savart Law

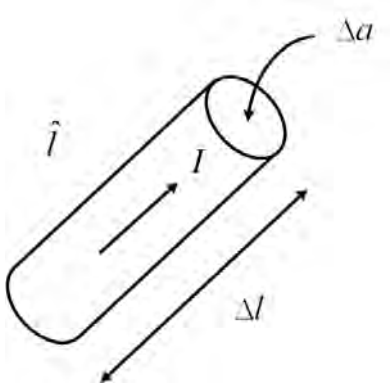


Figure 5.1: A current element used to illustrate the derivation of Biot-Savart law. The current element generates a magnetic field due to Ampere's law in the static limit. This law was established experimentally, but here, we will derive this law based on our mathematical knowledge so far.

Bio-Savart law allows us to derive the magnetic field due to the electric current flowing in a filamental wire. We assume that the wire is long, or that it forms a closed loop so that the current is uniform on the wire. To this end, we break the wire into union of tiny segments, and calculate the magnetic field from each of these tiny segments. From Gauss' law and Ampere's law in the static limit, and using the definition of the Green's function, we have derived that

$$\mathbf{A}(\mathbf{r}) = \frac{\mu}{4\pi} \iiint_V \frac{\mathbf{J}(\mathbf{r}')}{R} dV' \tag{5.1.1}$$

where $R = |\mathbf{r} - \mathbf{r}'|$, $\mathbf{r}$ is the field point (observation point), and $\mathbf{r}'$ is the source point. The above is just a three-dimensional convolutional integral between the Green's function and the source $\mathbf{J}(\mathbf{r}')$.

When the current element is small, the current is uniform across its cross section, and is carried by a wire of cross sectional area Δa as shown in Figure 5.1, we can approximate the integrand as

$$\mathbf{J}(\mathbf{r}')dV' \approx \mathbf{J}(\mathbf{r}')\Delta V' = \underbrace{\hat{l}I/\Delta a}_{\mathbf{J}(\mathbf{r}')} \underbrace{(\Delta a)\Delta l'}_{\Delta V'} = \hat{l}I\Delta l' \tag{5.1.2}$$

In the above, $\hat{l}I/\Delta a = \mathbf{J}(\mathbf{r}')$ is the current density, and $\Delta V = (\Delta a)\Delta l$ since $\mathbf{J}$ has the unit of amperes/m^2. Here, $\hat{l}$ is a unit vector pointing in the direction of the current flow or the axis of the wire. Hence, we can let the current element be

$$\mathbf{J}(\mathbf{r}')\Delta V' = I\Delta \mathbf{l}' \tag{5.1.3}$$

where the vector $\Delta\mathbf{l}' = \Delta l\hat{l}$, and $'$ indicates that it is located at $\mathbf{r}'$. Therefore, the incremental vector potential due to an incremental current element $\mathbf{J}(\mathbf{r}')\Delta V'$, according to (5.1.1), is

$$\Delta\mathbf{A}(\mathbf{r}) = \frac{\mu}{4\pi}\left(\frac{\mathbf{J}(\mathbf{r}')\Delta V'}{R}\right) = \frac{\mu}{4\pi}\frac{I\Delta\mathbf{l}'}{R} \tag{5.1.4}$$

Since $\mathbf{B} = \nabla\times\mathbf{A}$, we derive that this incremental $\mathbf{B}$ flux, $\Delta\mathbf{B}$ due to the incremental current $I\Delta\mathbf{l}'$, is

$$\Delta\mathbf{B} = \nabla\times\Delta\mathbf{A}(\mathbf{r}) = \frac{\mu I}{4\pi}\nabla\times\frac{\Delta\mathbf{l}'}{R} = \frac{-\mu I}{4\pi}\Delta\mathbf{l}'\times\nabla\frac{1}{R} \tag{5.1.5}$$

where ∇ operates on the field point, or the unprimed coordinates, and

$$R = \sqrt{(x-x')^2+(y-y')^2+(z-z')^2}$$

And we have made use of the fact that $\nabla\times\mathbf{a}f(\mathbf{r}) = -\mathbf{a}\times\nabla f(\mathbf{r})$ when $\mathbf{a}$ is a constant vector. The above can be simplified further by making use of the fact that[1]

$$\nabla\frac{1}{R} = -\frac{1}{R^2}\hat{R} \tag{5.1.6}$$

where $\hat{R}$ is a unit vector pointing in the $\mathbf{r}-\mathbf{r}'$ direction. Consequently, assuming that the incremental length is infinitesimally small, or $\Delta\mathbf{l}\to d\mathbf{l}$, we have, after using (5.1.6) in (5.1.5), that the incremental magnetic flux density $d\mathbf{B}$ is

$$\begin{aligned} d\mathbf{B} &= \frac{\mu I}{4\pi}d\mathbf{l}'\times\frac{1}{R^2}\hat{R} \\ &= \frac{\mu I d\mathbf{l}'\times\hat{R}}{4\pi R^2} \end{aligned} \tag{5.1.7}$$

Since $\mathbf{B} = \mu\mathbf{H}$, the incremental magnetic field is

$$d\mathbf{H} = \frac{Id\mathbf{l}'\times\hat{R}}{4\pi R^2} \tag{5.1.8}$$

or for contribution from the wire, assuming uniform current on the wire, is

$$\mathbf{H}(\mathbf{r}) = \int\frac{I(\mathbf{r}')d\mathbf{l}'\times\hat{R}}{4\pi R^2} \tag{5.1.9}$$

which is Biot-Savart law, first determined experimentally, but now derived using the rudiments of electromagnetic field theory.

[1] This is best done by expressing the ∇ operator and R in cartesian coordinates.

5.2 Shielding by Conductive Media

5.2.1 Boundary Conditions—Conductive Media Case

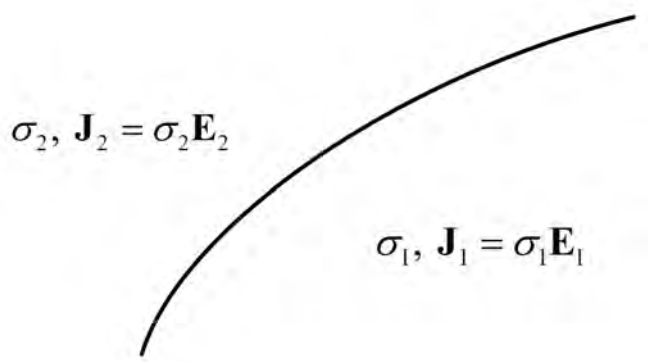


Figure 5.2: The schematics for deriving the boundary condition for the current density $\mathbf{J}$ at the interface of two conductive media.

In a conductive medium, $\mathbf{J} = \sigma \mathbf{E}$, which is just a statement of Ohm's law or $I = \frac{V}{R}$. From the current continuity equation, which is derivable from Ampere's law and Gauss' law for electric flux, one gets

$$\nabla \cdot \mathbf{J} = -\frac{\partial \varrho}{\partial t} \tag{5.2.1}$$

The above follows from the charge conservation law which cannot be violated. If the right-hand side is everywhere finite, it will not induce a jump discontinuity in the current. Moreover, it is zero for static limit. Hence, just like the Gauss's law case, the above implies that the normal component of the current J_n is continuous, or that $J_{1n} = J_{2n}$ in the static limit. In other words, in compact notation, the finiteness of $\nabla \cdot \mathbf{J}$ implies that

$$\hat{n} \cdot (\mathbf{J}_2 - \mathbf{J}_1) = 0 \tag{5.2.2}$$

Hence, using $\mathbf{J} = \sigma \mathbf{E}$, we have

$$\sigma_2 E_{2n} - \sigma_1 E_{1n} = 0 \tag{5.2.3}$$

Note that the above has to be always true in the static limit. But Gauss's law implies the boundary condition that

$$\varepsilon_2 E_{2n} - \varepsilon_1 E_{1n} = \varrho_s \tag{5.2.4}$$

The above equation (5.2.4) is incompatible with (5.2.3) unless $\varrho_s \neq 0$. Hence, surface charge density or charge accumulation is necessary at this interface, unless $\sigma_2/\sigma_1 = \varepsilon_2/\varepsilon_1$. This is found in semiconductor materials which are both conductive and having a permitivity: interfacial charges appear at the interface of two semi-conductor materials.

5.2.2 Electric Field Inside a Conductor

Dynamic Case:

For this case, the electric field inside a perfect electric conductor (PEC) has to be zero by the following argument. If medium 1 is a perfect electric conductor, then $\sigma \to \infty$ but $\mathbf{J}_1 = \sigma \mathbf{E}_1$. An infinitesimal $\mathbf{E}_1$ will give rise to an infinite current $\mathbf{J}_1$. To avoid this ludicrous situation, $\mathbf{E}_1$ has to be 0. This implies that $\mathbf{D}_1 = 0$ as well.

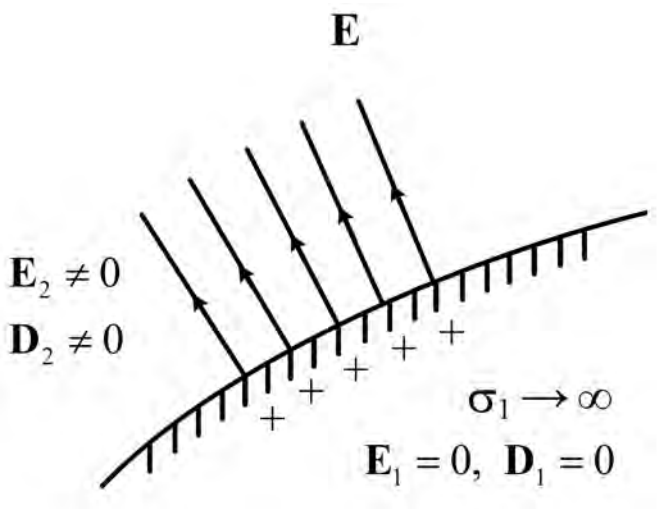


Figure 5.3: The behavior of the electric field and electric flux at the interface of a perfect electric conductor and free space (or air).

Since tangential $\mathbf{E}$ is continuous, from Faraday's law, it is still true that

$$E_{2t} = E_{1t} = 0 \tag{5.2.5}$$

or $\hat{n} \times \mathbf{E} = 0$. But since

$$\hat{n} \cdot (\mathbf{D}_2 - \mathbf{D}_1) = \varrho_s \tag{5.2.6}$$

and that $\mathbf{D}_1 = 0$, then

$$\hat{n} \cdot \mathbf{D}_2 = \varrho_s \tag{5.2.7}$$

So surface charge has to exist at a PEC/air interface. Moreover, normal $\mathbf{D}_2 \neq 0$, but tangential $\mathbf{E}_2 = 0$: Thus the $\mathbf{E}$ and $\mathbf{D}$ have to be normal to the PEC surface. The sketch of the electric field in the vicinity of a perfect electric conducting (PEC) surface is shown in Figure 5.3.

Static Case:

The above argument for zero electric field inside a perfect conductor is true for electrodynamic problems. However, one does not need the above argument regarding the shielding of the static electric field from a conducting region or an imperfect conductor. In the situation of the two conducting objects example below, as long as the electric fields are non-zero in the objects, currents will keep flowing. They will flow until the charges in the two objects orient themselves so that electric current cannot flow anymore. This happens when the charges produce internal fields that cancel each other giving rise to zero field inside the two objects. Faraday's law still applies which means that tangental **E** field has to be continuous but it is zero inside an imperfect conductor. Therefore, the boundary condition that the fields have to be normal to the conducting object surface is still true for electrostatics even if the conductor is imperfect. A sketch of the electric field between two conducting spheres is show in Figure 5.4. That is also the reason why electric charge can shield off electromagnetic radiation when the frequency is low as in a Faraday cage (see Figure 5.5).

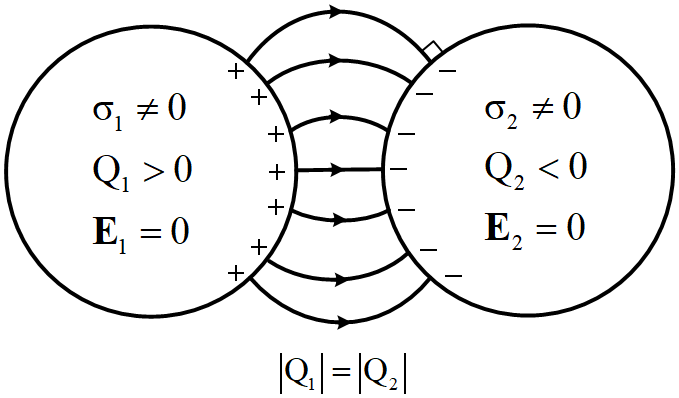


Figure 5.4: The behavior of the electric field and flux outside two conductors in the static limit. The two conductors need not be PEC, and yet, the fields are normal to the interface, and the fields are zero inside them.

Figure 5.5: (Left) The Faraday cage can be used to shield out low-frequency electromagnetic fields using charges on the surface of the cage, e.g., it is immune to lightning strike (courtesy of defendershield.com). (Right) It can be used to establish an electromagnetically quiet zone free from RF signals (courtesy of Wikipedia).

5.2.3 Magnetic Field Inside a Conductor

We have seen that for a finite conductor, as long as $\sigma \neq 0$, the charges will re-orient themselves until the electric field is expelled from the conductor; otherwise, the current will keep flowing until $\mathbf{E} = 0$. In a word, static $\mathbf{E}$ is zero inside a conductor.

But for the shielding of the magnetic field, the physics is different. There are no magnetic charges nor magnetic conductors in this world. Thus, this physical phenomenon for electric field does not happen for magnetic field: In other words, static magnetic field cannot be expelled from an imperfect electric conductor. However, a magnetic field can be expelled from a perfect conductor or a superconductor. You can only fully understand this physical phenomenon if we study the time-varying form of Maxwell's equations.

In a perfect conductor where $\sigma \to \infty$, it is unstable for the magnetic field $\mathbf{B}$ to be nonzero. As time varying magnetic field gives rise to an electric field by the time-varying form of Faraday's law, a small time variation of the $\mathbf{B}$ field will give rise to infinite current flow in a perfect conductor. Therefore to avoid this ludicrous situation, and to be stable, $\mathbf{B} = 0$ in a perfect conductor or a superconductor.

So if medium 1 is a perfect electric conductor (PEC), then $\mathbf{B}_1 = \mathbf{H}_1 = 0$. The boundary condition then for magnetic field from Ampere's law

$$\hat{n} \times (\mathbf{H}_2 - \mathbf{H}_1) = \hat{n} \times \mathbf{H}_2 = \mathbf{J}_s \tag{5.2.8}$$

which is the jump condition for the magnetic field. In other words, a surface current $\mathbf{J}_s$ has to flow at the surface of a PEC in order to support the jump discontinuity in the tangential component of the magnetic field.

From Gauss's law, $\nabla \cdot \mathbf{B} = 0$ implies that $\hat{n} \cdot \mathbf{B}$ is always continuous, or $\hat{n} \cdot (\mathbf{B}_2 - \mathbf{B}_1) = 0$, at an interface because of the absence of magnetic charges. But the magnetic flux $\mathbf{B}_1$ is expelled from the perfect conductor making $\hat{n} \cdot \mathbf{B}_1 = 0$ zero. Therefore, $\hat{n} \cdot \mathbf{B}_2 = 0$ as well. And hence, there is no

normal component of the **B** field at the interface of a PEC. Consequently, the boundary condition for $\mathbf{B}_2$ becomes, for a PEC,

$$\hat{n} \cdot \mathbf{B}_2 = 0 \tag{5.2.9}$$

The **B** field in the vicinity of a perfect conductor surface is as shown in Figure 5.6.

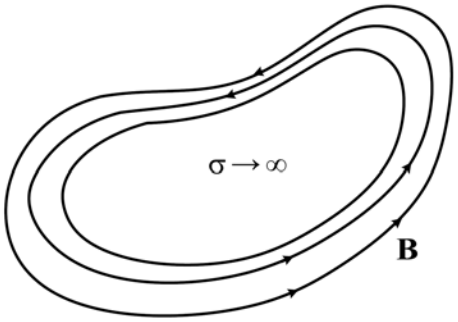


Figure 5.6: Sketch of the magnetic flux **B** around a perfect electric conductor (PEC). As explained in the text, it is seen that $\hat{n} \cdot \mathbf{B} = 0$ at the surface of the perfect electric conductor.

When a superconductor cube is placed next to a static magnetic field near a permanent magnet, the **B** has to be zero inside a superconductor due to the stability issue. Thus, eddy current will be induced on the superconductor to expel the magnetic field from the permanent magnet, or in a word, it will produce a magnetic dipole on the superconducting cube that repels the static magnetic field. Since these two magnetic dipoles are of opposite polarity, they repel each other. The repulsive force between the magnetic dipoles causes the superconducting cube to levitate on the static magnetic field as shown in Figure 5.7.[2]

[2]You may see this demo in a local science museum or your university's physics lab. I saw one in the Boston Museum of Science in 2018.

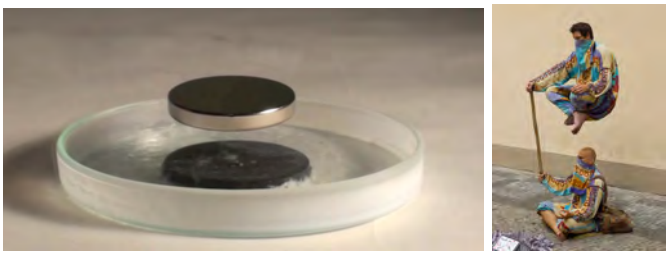


Figure 5.7: (Left) Levitation of a superconducting disk on top of a static magnetic field due to expulsion of the magnetic field from the superconductor. This is also known as the Meissner effect (figure courtesy of Wikimedia). (Right) Levitation by conjurors in a street in Prague. This effect is fake.

5.3 Instantaneous Poynting's Theorem

It is habitual to add fictitious magnetic current $\mathbf{M}$ and fictitious magnetic charge ϱ_m to Maxwell's equations to make them symmetric mathematically.[3] To this end, we have in general, Maxwell's equations as

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} - \mathbf{M} \tag{5.3.1}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J} \tag{5.3.2}$$

$$\nabla \cdot \mathbf{D} = \varrho \tag{5.3.3}$$

$$\nabla \cdot \mathbf{B} = \varrho_m \tag{5.3.4}$$

Before we proceed further with studying energy and power, Consider the first two of Maxwell's equations where fictitious magnetic current is included and that the medium is simple isotropic such that $\mathbf{B} = \mu_0 \mathbf{H}$ and $\mathbf{D} = \varepsilon_0 \mathbf{E}$. Next, we need to consider only the first two equations (since in electrodynamics, by invoking charge conservation, the third and the fourth equations are derivable from the first two). They are

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} - \mathbf{M}_i = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} - \mathbf{M}_i \tag{5.3.5}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J} = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J}_i + \sigma \mathbf{E} \tag{5.3.6}$$

where $\mathbf{M}_i$ and $\mathbf{J}_i$ are impressed current sources. They are sources that are impressed into the system, and they cannot be changed by their interaction with the environment [53]. They are

[3] Even though magnetic current does not exist, electric current can be engineered to look like magnetic current as shall be learnt. James Clerk Maxwell also added fictitious magnetic current in his mathematical treatise. Maxwell's equations are almost symmetric and it is tempting to contemplate the addition of the fictitious magnetic current to symmetrize them.

like current or voltage sources in circuit theory which are immutable by the environment they are inserted in.

Also, for a conductive medium, a conduction current or induced current flows in addition to impressed current. Here, $\mathbf{J} = \sigma\mathbf{E}$ is the induced current source in the conductor. Moreover, $\mathbf{J} = \sigma\mathbf{E}$ is similar to ohm's law. Maxwell's equations are mathematically elegant, and hence, they are amenable to manipulations by invoking mathematical logic, as we shall see.

By dot multiplying (5.3.5) with $\mathbf{H}$, and also dot multiplying (5.3.6) with $\mathbf{E}$, we can show that

$$\mathbf{H} \cdot \nabla \times \mathbf{E} = -\mu_0 \mathbf{H} \cdot \frac{\partial \mathbf{H}}{\partial t} - \mathbf{H} \cdot \mathbf{M}_i \tag{5.3.7}$$

$$\mathbf{E} \cdot \nabla \times \mathbf{H} = \varepsilon_0 \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} + \mathbf{E} \cdot \mathbf{J}_i + \sigma \mathbf{E} \cdot \mathbf{E} \tag{5.3.8}$$

Using the identity, which is the same as the product rule for derivatives, we have[4]

$$\nabla \cdot (\mathbf{E} \times \mathbf{H}) = \mathbf{H} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot (\nabla \times \mathbf{H}) \tag{5.3.9}$$

Therefore, from (5.3.7), (5.3.8), and (5.3.9), we have

$$\nabla \cdot (\mathbf{E} \times \mathbf{H}) = -\left(\mu_0 \mathbf{H} \cdot \frac{\partial \mathbf{H}}{\partial t} + \varepsilon_0 \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} + \sigma \mathbf{E} \cdot \mathbf{E} + \mathbf{H} \cdot \mathbf{M}_i + \mathbf{E} \cdot \mathbf{J}_i \right) \tag{5.3.10}$$

To elucidate the physical meaning of the above, we first consider $\sigma = 0$, and $\mathbf{M}_i = \mathbf{J}_i = 0$, or in the absence of conductive loss and the impressed current sources. Then the above becomes

$$\nabla \cdot (\mathbf{E} \times \mathbf{H}) = -\left(\mu_0 \mathbf{H} \cdot \frac{\partial \mathbf{H}}{\partial t} + \varepsilon_0 \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} \right) \tag{5.3.11}$$

Rewriting each term on the right-hand side of the above, we have[5]

$$\mu_0 \mathbf{H} \cdot \frac{\partial \mathbf{H}}{\partial t} = \frac{1}{2}\mu_0 \frac{\partial}{\partial t}(\mathbf{H} \cdot \mathbf{H}) = \frac{\partial}{\partial t}\left(\frac{1}{2}\mu_0 |\mathbf{H}|^2 \right) = \frac{\partial}{\partial t} W_m \tag{5.3.12}$$

$$\varepsilon_0 \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} = \frac{1}{2}\varepsilon_0 \frac{\partial}{\partial t}(\mathbf{E} \cdot \mathbf{E}) = \frac{\partial}{\partial t}\left(\frac{1}{2}\varepsilon_0 |\mathbf{E}|^2 \right) = \frac{\partial}{\partial t} W_e \tag{5.3.13}$$

where $|\mathbf{H}(\mathbf{r}, t)|^2 = \mathbf{H}(\mathbf{r}, t) \cdot \mathbf{H}(\mathbf{r}, t)$, and $|\mathbf{E}(\mathbf{r}, t)|^2 = \mathbf{E}(\mathbf{r}, t) \cdot \mathbf{E}(\mathbf{r}, t)$ are positive definite. Then (5.3.11) becomes

$$\nabla \cdot (\mathbf{E} \times \mathbf{H}) = -\frac{\partial}{\partial t}\left(W_m + W_e \right) \tag{5.3.14}$$

where

$$W_m = \frac{1}{2}\mu_0 |\mathbf{H}|^2, \qquad W_e = \frac{1}{2}\varepsilon_0 |\mathbf{E}|^2 \tag{5.3.15}$$

[4]The cyclical identity, or the cyclical triple product rule, that $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a})$ is useful for the derivation.

[5]The following equality can be established by the product rule of differentiation that $\frac{\partial}{\partial t}(\mathbf{H} \cdot \mathbf{H}) = \mathbf{H} \cdot \frac{\partial \mathbf{H}}{\partial t} + \frac{\partial \mathbf{H}}{\partial t} \cdot \mathbf{H}$.

The above are energy densities in the magnetic field and electric field, respectively.

The vector quantity

$$\mathbf{S}_p = \mathbf{E} \times \mathbf{H} \tag{5.3.16}$$

is called the Poynting's vector, and (5.3.14) becomes

$$\nabla \cdot \mathbf{S}_p = -\frac{\partial}{\partial t} W_t \tag{5.3.17}$$

where $W_t = W_e + W_m$ is the total energy density stored in the electric and magnetic fields while $\mathbf{S}_p$ is the power density. It is easy to show that $\mathbf{S}_p$, the power density, has a dimension of watts per meter square, and that W_t, the energy density, has a dimension of joules per meter cube.

The above is similar in physical interpretation to the current continuity equation which is

$$\nabla \cdot \mathbf{J}(\mathbf{r}, t) = -\partial_t \varrho(\mathbf{r}, t) \tag{5.3.18}$$

One can think that in the current continuity equation, that current density is charge density flow. Hence, power density is energy density flow. We can think of a cube of energy density W_t moving at velocity v, giving rise to power density S_p, and their relationship is

$$S_p = W_t v \tag{5.3.19}$$

The right-hand side represents energy density flow while the left-hand side represents power density. One can check the sanity of the above equation using dimensional analysis.

Now, if we let $\sigma \neq 0$, then the term to be included is then $\sigma \mathbf{E} \cdot \mathbf{E} = \sigma |\mathbf{E}|^2$ which has the unit of (S m^{-1}) times (V^2 m^{-2}), or (W m^{-3}) where S is siemens. (We arrive at this unit by noticing that $\frac{1}{2}\frac{V^2}{R}$ is the power dissipated in a resistor of R ohms with a unit of watts.) The reciprocal unit of ohms, which used to be called mhos is now called siemens. With $\sigma \neq 0$, (5.3.17) becomes

$$\nabla \cdot \mathbf{S}_p = -\frac{\partial}{\partial t} W_t - \sigma |\mathbf{E}|^2 = -\frac{\partial}{\partial t} W_t - P_d \tag{5.3.20}$$

Here, $\nabla \cdot \mathbf{S}_p$ has the physical meaning of power density oozing out from a point, and $-P_d = -\sigma |\mathbf{E}|^2$ has the physical meaning of power density dissipated (siphoned) at a point by the conductive loss in the medium which is proportional to $-\sigma |\mathbf{E}|^2$.

Now if we set $\mathbf{J}_i$ and $\mathbf{M}_i$ to be nonzero, (5.3.20) is augmented by the last two terms in (5.3.10), or

$$\nabla \cdot \mathbf{S}_p = -\frac{\partial}{\partial t} W_t - P_d - \mathbf{H} \cdot \mathbf{M}_i - \mathbf{E} \cdot \mathbf{J}_i \tag{5.3.21}$$

The last two terms can be interpreted as the power density supplied by the impressed currents $\mathbf{M}_i$ and $\mathbf{J}_i$ or in short power source P_s. Therefore, (5.3.21) becomes

$$\nabla \cdot \mathbf{S}_p = -\frac{\partial}{\partial t} W_t - P_d + P_s \tag{5.3.22}$$

where

$$P_s = -\mathbf{H} \cdot \mathbf{M}_i - \mathbf{E} \cdot \mathbf{J}_i \tag{5.3.23}$$

and P_s is the power supplied by the impressed current sources $\mathbf{M}_i$ and $\mathbf{J}_i$. These terms are positive if $\mathbf{H}$ and $\mathbf{M}_i$ have opposite signs, and if $\mathbf{E}$ and $\mathbf{J}_i$ have opposite signs. The last terms reminds us of what happens in a negative resistance device or in a battery.[6] In a battery, positive charges move from a region of lower potential to a region of higher potential (see Figure 5.8) as opppose to those in a resistor. The positive charges move from one end of a battery to the other end of the battery. Hence, they are doing an "uphill climb" driven by chemical processes within the battery.

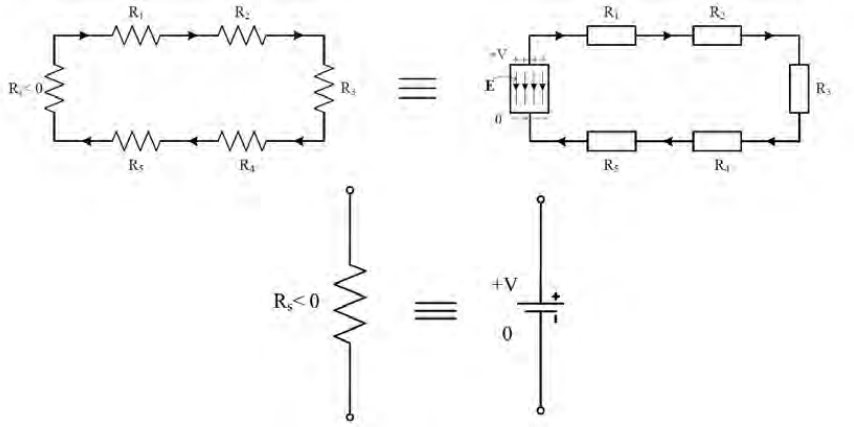


Figure 5.8: Figure showing the dissipation of energy as the current flows around a loop. A battery can be viewed as having a negative resistance.

In the above, one can easily work out that P_s has the unit of W m^{-3} which is power density supplied. One can also choose to rewrite (5.3.22) in integral form by integrating it over a volume V and invoking the divergence theorem yielding

$$\int_S d\mathbf{S} \cdot \mathbf{S}_p = -\frac{d}{dt} \int_V W_t dV - \int_V P_d dV + \int_V P_s dV \tag{5.3.24}$$

The left-hand side is

$$\int_S d\mathbf{S} \cdot \mathbf{S}_p = \int_S d\mathbf{S} \cdot (\mathbf{E} \times \mathbf{H}) \tag{5.3.25}$$

which represents the power flowing out of the surface S.

[6] A negative resistance has been made by Leo Esaki [54] in an electronic tunnel diode, winning him a share in the Nobel prize.

Exercises for Lecture 5

Problem 5-1:

(i) Derive eqs. (5.1.5) and (5.1.6).

(ii) Explain why for electrostatics, perfect conductor is not needed to shield out the electric field.

(iii) Explain the Meissner effect in a superconductor, and why a small piece of superconductor can levitate on a pole of a permanent magnet.

(iv) Give physical interpretation to equation (5.3.20) and the meaning of each of the terms in the equation.

Chapter 6

Time-Harmonic Fields, Complex Power

The analysis of Maxwell's equations can be greatly simplified by assuming the fields to be time harmonic, or sinusoidal (cosinusoidal).[1] Electrical engineers use a method called phasor technique [34, 55], to simplify equations involving time-harmonic signals (they are variously known as monochromatic or CW (constant wave) signals). This is also a poor-man's Fourier transform [56]. That is, one begets the benefits of Fourier transform technique without the full knowledge of Fourier transform. Since only one time-harmonic frequency is involved, this is also called frequency domain analysis, or that only one frequency component of the Fourier transform of the signal is involved. Phasors are represented in complex numbers. Therefore, the fields become complex in the frequency domain. From this, we will also discuss the concept of complex power.

[1] It is simple only for linear systems: for nonlinear systems, such analysis can be quite unwieldy. But rest assured, as we will not discuss nonlinear systems in this course.

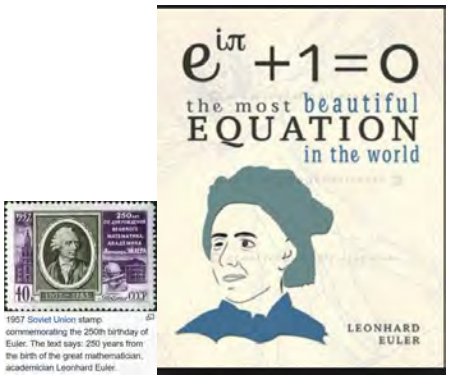


Figure 6.1: A commemorative stamp showing the contribution of Euler (courtesy of Wikipedia and Pinterest).

6.1 Time-Harmonic Fields—Linear Systems

To learn phasor technique, one makes use the formula due to Euler (1707–1783) (Wikipedia)[2]

$$e^{j\alpha} = \cos\alpha + j\sin\alpha \tag{6.1.1}$$

where $j = \sqrt{-1}$ is an imaginary number.[3]

From Euler's formula, one gets

$$\cos\alpha = \Re e\left(e^{j\alpha}\right) \tag{6.1.2}$$

where $\Re e$ stands for "the real part of". Hence, all time harmonic quantities can be written as

$$\begin{aligned} V(x,y,z,t) &= V'(x,y,z)\cos(\omega t+\alpha) && (6.1.3)\\ &= V'(\mathbf{r})\Re e(e^{j(\omega t+\alpha)}) && (6.1.4)\\ &= \Re e\left(V'(\mathbf{r})e^{j\alpha}e^{j\omega t}\right) && (6.1.5)\\ &= \Re e\left(\underset{\sim}{V}(\mathbf{r})e^{j\omega t}\right) && (6.1.6) \end{aligned}$$

Now $\underset{\sim}{V}(\mathbf{r}) = V'(\mathbf{r})e^{j\alpha}$ is a complex number called the phasor representation or phasor of $V(\mathbf{r},t)$, a time-harmonic quantity.[4] Here, the phase $\alpha = \alpha(\mathbf{r})$ can also be a function of position $\mathbf{r}$, or x, y, z.

[2] As the stamp shows, Euler was blind in one eye.

[3] But lo and behold, in other disciplines such as physics, mathematics, and optics, $\sqrt{-1}$ is denoted by "i", but "i" is too close to the symbol for current. So the preferred symbol for electrical engineering for an imaginary number is j: a quirkness of convention, just as positive charges do not carry current in a wire.

[4] We will use under tilde to denote a complex number or a phasor here, but this notation will be dropped later. Whether a variable is complex or real is clear from the context.

Consequently, any component of a field can be expressed as[5]

$$E_x(x,y,z,t) = E_x(\mathbf{r},t) = \Re e\left[\underset{\sim}{E}_x(\mathbf{r})e^{j\omega t}\right] \tag{6.1.7}$$

The above can be repeated for y and z components. Compactly, for the x, y, and z components together, one can write

$$\mathbf{E}(\mathbf{r},t) = \Re e\left[\underset{\sim}{\mathbf{E}}(\mathbf{r})e^{j\omega t}\right] \tag{6.1.8}$$

$$\mathbf{H}(\mathbf{r},t) = \Re e\left[\underset{\sim}{\mathbf{H}}(\mathbf{r})e^{j\omega t}\right] \tag{6.1.9}$$

where $\underset{\sim}{\mathbf{E}}$ and $\underset{\sim}{\mathbf{H}}$ are complex vector fields. It is to be noted that the phasor representation of a field is a complex number but it has the same dimension (or unit) as the real field.

Such phasor representations of time-harmonic fields simplify Maxwell's equations. For instance, if one writes

$$\mathbf{B}(\mathbf{r},t) = \Re e\left[\underset{\sim}{\mathbf{B}}(\mathbf{r})e^{j\omega t}\right] \tag{6.1.10}$$

then

$$\begin{aligned}\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r},t) &= \frac{\partial}{\partial t}\Re e\left[\underset{\sim}{\mathbf{B}}(\mathbf{r})e^{j\omega t}\right]\\ &= \Re e\left[\frac{\partial}{\partial t}\underset{\sim}{\mathbf{B}}(\mathbf{r})e^{j\omega t}\right]\\ &= \Re e\left[\underset{\sim}{\mathbf{B}}(\mathbf{r})j\omega e^{j\omega t}\right]\end{aligned} \tag{6.1.11}$$

Therefore, a time derivative can be effected very simply for a time-harmonic field: One just needs to multiply $j\omega$ to the phasor representation of a field or a signal. Hence, given Faraday's law that

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} - \mathbf{M} \tag{6.1.12}$$

assuming that all quantities are time harmonic, then with (6.1.10) and what follows,

$$\mathbf{E}(\mathbf{r},t) = \Re e\left[\underset{\sim}{\mathbf{E}}(\mathbf{r})e^{j\omega t}\right] \tag{6.1.13}$$

$$\mathbf{M}(\mathbf{r},t) = \Re e\left[\underset{\sim\sim}{\mathbf{M}}(\mathbf{r})e^{j\omega t}\right] \tag{6.1.14}$$

using (6.1.11) and the above into (6.1.12), one gets first

$$\nabla\times\mathbf{E}(\mathbf{r},t) = \Re e\left[\nabla\times\underset{\sim}{\mathbf{E}}(\mathbf{r})e^{j\omega t}\right] \tag{6.1.15}$$

[5] In some area, this is often written as $E_x(\mathbf{r},t) = \frac{1}{2}\underset{\sim}{E}_x(\mathbf{r})e^{j\omega t} + c.c.$ where "c.c." stands for "complex conjugate" or "hermitian conjugate".

and then equating to the right-hand side of (6.1.12), we have

$$\Re e\left[\nabla \times \underset{\sim}{\mathbf{E}}(\mathbf{r})e^{j\omega t}\right] = -\Re e\left[\underset{\sim}{\mathbf{B}}(\mathbf{r})j\omega e^{j\omega t}\right] - \Re e\left[\underset{\sim}{\mathbf{M}}(\mathbf{r})e^{j\omega t}\right] \tag{6.1.16}$$

Since if

$$\Re e\left[A(\mathbf{r})e^{j\omega t}\right] = \Re e\left[B(\mathbf{r})e^{j\omega t}\right], \quad \forall t \tag{6.1.17}$$

then $A(\mathbf{r}) = B(\mathbf{r})$; it must be true from (6.1.16) that[6]

$$\nabla \times \underset{\sim}{\mathbf{E}}(\mathbf{r}) = -j\omega \underset{\sim}{\mathbf{B}}(\mathbf{r}) - \underset{\sim}{\mathbf{M}}(\mathbf{r}) \tag{6.1.18}$$

Therefore, finding the phasor representation of an equation in the frequency domain is clear: whenever we have $\frac{\partial}{\partial t}$, we replace it by $j\omega$. Applying this methodically to the other Maxwell's equations, we have

$$\nabla \times \underset{\sim}{\mathbf{H}}(\mathbf{r}) = j\omega \underset{\sim}{\mathbf{D}}(\mathbf{r}) + \underset{\sim}{\mathbf{J}}(\mathbf{r}) \tag{6.1.19}$$

$$\nabla \cdot \underset{\sim}{\mathbf{D}}(\mathbf{r}) = \underset{\sim}{\varrho}_e(\mathbf{r}) \tag{6.1.20}$$

$$\nabla \cdot \underset{\sim}{\mathbf{B}}(\mathbf{r}) = \underset{\sim}{\varrho}_m(\mathbf{r}) \tag{6.1.21}$$

In the above, the phasors are functions of frequency. For instance, $\underset{\sim}{\mathbf{H}}(\mathbf{r})$ should rightly be written as $\underset{\sim}{\mathbf{H}}(\mathbf{r}, \omega)$, but the ω dependence is implied.

6.2 Fourier Transform Technique

An advantage of phasor technique is that the phasor representations of the fields have the same unit as the original fields in the time domain, as can be seen (6.1.8) and (6.1.9). This is not the case for Fourier transform technique.

In the phasor representation, Maxwell's equations has no time derivatives; hence, the equations are simplified. We can also arrive at the above simplified Maxwell's equations using Fourier transform technique. To this end, we use Faraday's law as an example. By letting

$$\mathbf{E}(\mathbf{r}, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{E}(\mathbf{r}, \omega)e^{j\omega t} d\omega \tag{6.2.1}$$

$$\mathbf{B}(\mathbf{r}, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{B}(\mathbf{r}, \omega)e^{j\omega t} d\omega \tag{6.2.2}$$

$$\mathbf{M}(\mathbf{r}, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{M}(\mathbf{r}, \omega)e^{j\omega t} d\omega \tag{6.2.3}$$

[6]The converse is definitely true. If the forward is false, and the converse is true, that will be absurd, and hence, the forward is true by *reducatio ad absurdum*.

Substituting the above into Faraday's law given by (6.1.12), we get

$$\nabla \times \int_{-\infty}^{\infty} d\omega e^{j\omega t}\mathbf{E}(\mathbf{r},\omega) = -\frac{\partial}{\partial t}\int_{-\infty}^{\infty} d\omega e^{j\omega t}\mathbf{B}(\mathbf{r},\omega) - \int_{-\infty}^{\infty} d\omega e^{j\omega t}\mathbf{M}(\mathbf{r},\omega) \tag{6.2.4}$$

Using the fact that

$$\frac{\partial}{\partial t}\int_{-\infty}^{\infty} d\omega e^{j\omega t}\mathbf{B}(\mathbf{r},\omega) = \int_{-\infty}^{\infty} d\omega \frac{\partial}{\partial t} e^{j\omega t}\mathbf{B}(\mathbf{r},\omega) = \int_{-\infty}^{\infty} d\omega e^{j\omega t} j\omega \mathbf{B}(\mathbf{r},\omega) \tag{6.2.5}$$

and by exchanging the order of differentiation and integration, that

$$\nabla \times \int_{-\infty}^{\infty} d\omega e^{j\omega t}\mathbf{E}(\mathbf{r},\omega) = \int_{-\infty}^{\infty} d\omega e^{j\omega t}\nabla \times \mathbf{E}(\mathbf{r},\omega) \tag{6.2.6}$$

Furthermore, using the fact that

$$\int_{-\infty}^{\infty} d\omega e^{j\omega t} A(\omega) = \int_{-\infty}^{\infty} d\omega e^{j\omega t} B(\omega), \quad \forall t \tag{6.2.7}$$

implies that $A(\omega) = B(\omega)$, and using (6.2.5) and (6.2.6) in (6.2.4), and the property (6.2.7), one gets

$$\nabla \times \mathbf{E}(\mathbf{r},\omega) = -j\omega \mathbf{B}(\mathbf{r},\omega) - \mathbf{M}(\mathbf{r},\omega) \tag{6.2.8}$$

These equations look exactly like the phasor equations we have derived previously, save that the field $\mathbf{E}(\mathbf{r},\omega)$, $\mathbf{B}(\mathbf{r},\omega)$, and $\mathbf{M}(\mathbf{r},\omega)$ are now the Fourier transforms of the field $\mathbf{E}(\mathbf{r},t)$, $\mathbf{B}(\mathbf{r},t)$, and $\mathbf{M}(\mathbf{r},t)$. Moreover, the Fourier transform variables can be complex just like phasors. Repeating the exercise above for the other Maxwell's equations, we obtain equations that look similar to those for their phasor representations. Hence, Maxwell's equations can be simplified either by using phasor technique or Fourier transform technique. However, the dimensions (or units) of the phasors are different from the dimensions of the Fourier-transformed fields: $\underset{\sim}{\mathbf{E}}(\mathbf{r})$, a phasor, and $\mathbf{E}(\mathbf{r},\omega)$, a Fourier transform, do not have the same dimension upon closer examination. The advantage of phasor technique is that the units of the fields are not changed. Also, phasor quantities are related to Fourier transform quantities just by a multiplicative constant.

6.3 Complex Power

Consider now that in the phasor representations, $\underset{\sim}{\mathbf{E}}(\mathbf{r})$ and $\underset{\sim}{\mathbf{H}}(\mathbf{r})$ are complex vectors, and their cross product, $\underset{\sim}{\mathbf{E}}(\mathbf{r}) \times \underset{\sim}{\mathbf{H}}^*(\mathbf{r})$, which still has the unit of power density, has a different physical meaning. First, consider the instantaneous Poynting's vector

$$\mathbf{S}(\mathbf{r},t) = \mathbf{E}(\mathbf{r},t) \times \mathbf{H}(\mathbf{r},t) \tag{6.3.1}$$

where all the quantities are real valued but $\mathbf{E}$ and $\mathbf{H}$ are time-harmonic. Now, we can use phasor technique to analyze the above. Since they are time-harmonic fields, the above can be rewritten as

$$\begin{aligned}\mathbf{S}(\mathbf{r},t) &= \Re e\left[\underset{\sim}{\mathbf{E}}(\mathbf{r})e^{j\omega t}\right]\times\Re e\left[\underset{\sim}{\mathbf{H}}(\mathbf{r})e^{j\omega t}\right]\\ &= \frac{1}{2}\left[\underset{\sim}{\mathbf{E}}e^{j\omega t}+(\underset{\sim}{\mathbf{E}}e^{j\omega t})^*\right]\times\frac{1}{2}\left[\underset{\sim}{\mathbf{H}}e^{j\omega t}+(\underset{\sim}{\mathbf{H}}e^{j\omega t})^*\right]\end{aligned}\tag{6.3.2}$$

where we have made use of the formula that

$$\Re e(Z)=\frac{1}{2}(Z+Z^*)\tag{6.3.3}$$

Then more elaborately, on expanding (6.3.2), we get

$$\mathbf{S}(\mathbf{r},t)=\frac{1}{4}\underset{\sim}{\mathbf{E}}\times\underset{\sim}{\mathbf{H}}e^{2j\omega t}+\frac{1}{4}\underset{\sim}{\mathbf{E}}\times\underset{\sim}{\mathbf{H}}^*+\frac{1}{4}\underset{\sim}{\mathbf{E}}^*\times\underset{\sim}{\mathbf{H}}+\frac{1}{4}\underset{\sim}{\mathbf{E}}^*\times\underset{\sim}{\mathbf{H}}^*e^{-2j\omega t}\tag{6.3.4}$$

On rearranging terms and using (6.3.3) yield

$$\mathbf{S}(\mathbf{r},t)=\frac{1}{2}\Re e\left[\underset{\sim}{\mathbf{E}}\times\underset{\sim}{\mathbf{H}}^*\right]+\frac{1}{2}\Re e\left[\underset{\sim}{\mathbf{E}}\times\underset{\sim}{\mathbf{H}}e^{2j\omega t}\right]\tag{6.3.5}$$

where the first term is independent of time, while the second term is sinusoidal in time. If we define a time-average quantity such that

$$\mathbf{S}_{\text{av}}=\langle\mathbf{S}(\mathbf{r},t)\rangle=\lim_{T\to\infty}\frac{1}{T}\int_0^T\mathbf{S}(\mathbf{r},t)dt\tag{6.3.6}$$

it is quite clear now that the second term of (6.3.5) time-averages to zero since it is sinusoidal, and

$$\mathbf{S}_{\text{av}}=\langle\mathbf{S}(\mathbf{r},t)\rangle=\frac{1}{2}\Re e\left[\underset{\sim}{\mathbf{E}}\times\underset{\sim}{\mathbf{H}}^*\right]\tag{6.3.7}$$

Therefore, in the phasor representation, the quantity

$$\underset{\sim}{\mathbf{S}}=\underset{\sim}{\mathbf{E}}\times\underset{\sim}{\mathbf{H}}^*\tag{6.3.8}$$

is termed the complex Poynting's vector which has the dimension of watts per square meter. The complex power density $\underset{\sim}{\mathbf{S}}$, is energy density flow associated with it, and is associated with complex power.[7] We have understood what the real part of this complex power is, and next, we will elucidate the physical meaning of its imaginary part.

[7]Please note that $\underset{\sim}{\mathbf{S}}$ is not the phasor representation of $\mathbf{S}(\mathbf{r},t)$ since the latter is not a time-harmonic signal anymore, and has no phasor representation.

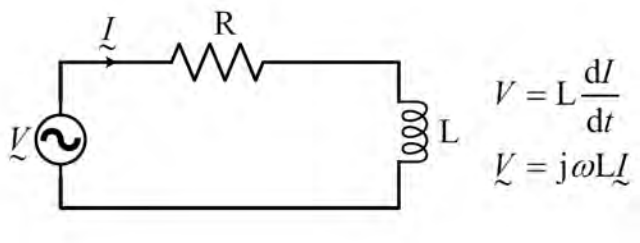


Figure 6.2: A simple circuit example to illustrate the concept of complex power in circuit theory. The voltage and current are out of phase which is a frequency-domain concept.

To understand what complex power is , it is fruitful if we revisit complex power [53, 57] in our circuit theory course. The circuit in Figure 6.2 can be easily solved by using phasor technique. The impedance of the circuit is $Z = R + j\omega L$. Hence,

$$\underset{\sim}{V} = (R + j\omega L)\underset{\sim}{I} \tag{6.3.9}$$

where $\underset{\sim}{V}$ and $\underset{\sim}{I}$ are the phasors of the voltage and current for time-harmonic signals. Just as in the electromagnetic case, the complex power in watts is taken to be

$$\underset{\sim}{P} = \underset{\sim}{V}\underset{\sim}{I}^* \tag{6.3.10}$$

But the instantaneous power is given by

$$P_{\text{inst}}(t) = V(t)I(t) \tag{6.3.11}$$

where $V(t) = \Re e\{\underset{\sim}{V}e^{j\omega t}\}$ and $I(t) = \Re e\{\underset{\sim}{I}e^{j\omega t}\}$. As shall be shown below,

$$P_{\text{av}} = \langle P_{\text{inst}}(t)\rangle = \frac{1}{2}\Re e\left[\underset{\sim}{P}\right] \tag{6.3.12}$$

It is clear that if $V(t)$ is sinusoidal, it can be written as

$$V(t) = V_0\cos(\omega t) = \Re e\left[\underset{\sim}{V}e^{j\omega t}\right] \tag{6.3.13}$$

where, without loss of generality, we assume that $\underset{\sim}{V} = V_0$. Then from (6.3.9), it is clear that $V(t)$ and $I(t)$ are not in phase. Namely that

$$I(t) = I_0\cos(\omega t + \alpha) = \Re e\left[\underset{\sim}{I}e^{j\omega t}\right] \tag{6.3.14}$$

where $\underset{\sim}{I} = I_0 e^{j\alpha}$. Then

$$\begin{aligned} P_{\text{inst}}(t) &= V_0 I_0\cos(\omega t)\cos(\omega t + \alpha) \\ &= V_0 I_0\cos(\omega t)\left[\cos(\omega t)\cos(\alpha) - \sin(\omega t)\sin\alpha\right] \\ &= V_0 I_0\cos^2(\omega t)\cos\alpha - V_0 I_0\cos(\omega t)\sin(\omega t)\sin\alpha \end{aligned} \tag{6.3.15}$$

It can be seen that the first term does not time-average to zero, but the second term does! (To see this, we let $\cos(\omega t)\sin(\omega t) = 0.5\sin(2\omega t)$, which time-average to zero.) Now taking the time average of (6.3.15), the time average of the first term involves the time average of $\cos^2(\omega t)$ which is 0.5, we get

$$P_{\rm av} = \langle P_{\rm inst}\rangle = \frac{1}{2}V_0 I_0\cos\alpha = \frac{1}{2}\Re e\left[\underset{\sim}{V}\underset{\sim}{I}^*\right] \tag{6.3.16}$$

$$= \frac{1}{2}\Re e\left[\underset{\sim}{P}\right] \tag{6.3.17}$$

On the other hand, since the power is now complex, the imagine part of P is called the reactive power. It has a physical meaning that we shall elucidate next.

$$P_{\rm reactive} = \frac{1}{2}\Im m\left[\underset{\sim}{P}\right] = \frac{1}{2}\Im m\left[\underset{\sim}{V}\underset{\sim}{I}^*\right] = \frac{1}{2}\Im m\left[V_0 I_0 e^{-j\alpha}\right] = -\frac{1}{2}V_0 I_0\sin\alpha \tag{6.3.18}$$

One sees that amplitude of the time-varying term in (6.3.15) is precisely proportional to $\Im m\left[\underset{\sim}{P}\right]$.[8]

The reason for the existence of the imaginary part of $\underset{\sim}{P}$ is because $V(t)$ and $I(t)$ are out of phase or $\underset{\sim}{V} = V_0$, but $\underset{\sim}{I} = I_0 e^{j\alpha}$. They are out of phase is because the impedance of the circuit shown with an inductor has a reactive part to it. Hence the imaginary part of complex power is also called the reactive power [53, 57, 37]. In a reactive circuit, the plots of the instantaneous power is shown in Figure 6.3. It is seen that when $\alpha \neq 0$, the instantaneous power can be negative at certain instant. This means at this instant, the power is flowing from the load to the source instead of flowing from the source to the load. This happens only when the reactive power is nonzero or when a reactive component like an inductor or capacitor exists in the circuit.

When a power company delivers power to our homes, the power is complex because the current and voltage are not in phase due to the presence of rotating machineries, computers, transformers etc in our home appliances. Even though the reactive power time-averages to zero, the power company still needs to deliver it to-and-fro our home to run our washing machine, dish washer, fans, and air conditioner etc. And hence, they charge us for it. To be fair, part of this power will be dissipated in the transmission lines that deliver power to-and-fro our homes. In other words, we have to pay for the use of imaginary power! One advantage of DC power sources is that there is no reactive power associated with them. The con against them is that they cannot be stepped or down in voltages easily.

[8] Because that complex power is proportional to $\underset{\sim}{V}\underset{\sim}{I}^*$, it is the relative phase between $\underset{\sim}{V}$ and $\underset{\sim}{I}$ that matters. Therefore, α above is the relative phase between the phasor current and phasor voltage.

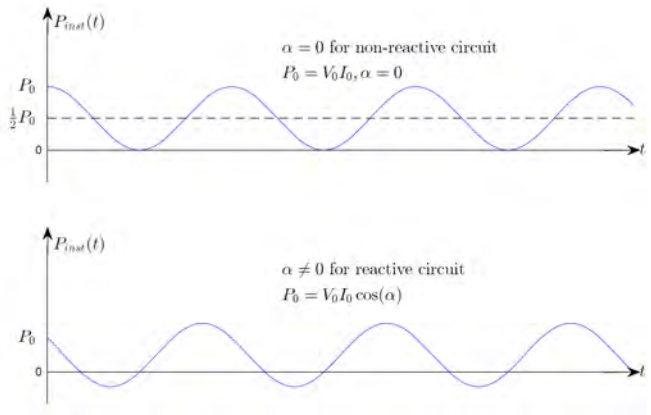


Figure 6.3: (Top) Plots of instantaneous power for when the voltage and the current is in phase ($\alpha = 0$), (Bottom) and when they are out of phase ($\alpha \neq 0$). In the out-of-phase case, there is an additional time-varying term that does not contribute to time-average power as shown in (6.3.15). But the instantaneous power can be negative as shown.

Exercises for Lecture 6

Problem 6-1:

(i) Explain why equation (6.1.17) and the statement after it is true.

(ii) Explain why equation (6.2.7) and the statement after it is true.

(iii) Is there a difference in the field quantities obtained from phasor technique and the field quantities obtained from Fourier transform technique?

(iv) Explain the physical meaning of the imaginary part of complex power.

Chapter 7

More on Constitutive Relations, Uniform Plane Wave

As mentioned before, for time-varying problems, the last two of Maxwell's equations are derivable from the first two. But constitutive relations are important for us to solve only the first two of four Maxwell's equations. Assuming that $\mathbf{J}$ is known or zero, then the first two vector equations have four vector unknowns: $\mathbf{E}$, $\mathbf{H}$, $\mathbf{D}$, and $\mathbf{B}$, which cannot be determined by solving only two equations. The additional two equations come from the constitutive relations. Constitutive relations are useful because they allow us to incorporate material properties into the solutions of Maxwell's equations. The material properties can be frequency dispersive (functions of frequency), anisotropic, bi-anisotropic, inhomogeneous, lossy, conductive, nonlinear as well as spatially dispersive. The use of phasors or frequency domain method will further simplify the characterization of different media. To begin, we will also study uniform plane wave in such media, including lossy conductive media.

7.1 More on Constitutive Relations

As have been said, Maxwell's equations are not solvable until the constitutive relations are included. Here, we will look more deeply into various kinds of constitutive relations. Now that we have learned phasor technique which is a powerful tool for frequency domain analysis, we can study a more general constitutive relationship compared to what we have seen earlier.

7.1.1 Isotropic Frequency Dispersive Media

First let us look at the simple linear constitutive relation previously discussed for dielectric media where [33], [34][p. 82], [49]

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} \tag{7.1.1}$$

We begin with a simple linear model where

$$\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E} \tag{7.1.2}$$

where χ_e is the electric susceptibility. Since displacement current is $\frac{\partial \mathbf{D}}{\partial t}$, when used in the generalized Ampere's law, $\mathbf{P}$, the polarization density, plays an important role for enhancing the flow of the displacement current through space. The polarization density is due to the presence of polar atoms or molecules that behave like little dipoles in the presence of an electric field. For instance, water, whose molecule is H_2O, is a polar molecule that becomes a small dipole when an electric field is applied.

We can think of displacement current flow in polar molecules as capacitive coupling between the dipoles yielding polarization current that flows through space. Namely, for a source-free medium,

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t} \tag{7.1.3}$$

Figure 7.1: As a series of dipoles lined up end to end, one can see a current, through capacitive coupling, flowing through the line of dipoles as they oscillate back and forth in their polarity. This is similar to how displacement current flows through a series of capacitors.

For example, for a sinusoidal oscillating field, the dipoles will flip back and forth giving rise to flow of displacement current just as how time-harmonic electric current can flow through a capacitor as shown in Figure 7.1.

The linear relationship above can be generalized to that of a linear time-invariant system [55], or that at any given space point $\mathbf{r}$ [37][p. 212], [49][p. 330].

$$\mathbf{P}(\mathbf{r}, t) = \varepsilon_0 \chi_e(\mathbf{r}, t) \circledast \mathbf{E}(\mathbf{r}, t) \tag{7.1.4}$$

where $\circledast$ here implies a convolution. In the frequency domain or the Fourier space, the above linear relationship becomes simpler which is

$$\mathbf{P}(\mathbf{r}, \omega) = \varepsilon_0 \chi_e(\mathbf{r}, \omega) \mathbf{E}(\mathbf{r}, \omega), \tag{7.1.5}$$

or

$$\mathbf{D}(\mathbf{r}, \omega) = \varepsilon_0 \left[1 + \chi_e(\mathbf{r}, \omega)\right] \mathbf{E}(\mathbf{r}, \omega) = \varepsilon(\mathbf{r}, \omega) \mathbf{E}(\mathbf{r}, \omega) \tag{7.1.6}$$

where $\varepsilon(\mathbf{r}, \omega) = \varepsilon_0 \left[1 + \chi_e(\mathbf{r}, \omega)\right]$ at any point $\mathbf{r}$ in space. There is a rich variety of ways at which $\chi_e(\omega)$ can manifest itself. Such a permittivity $\varepsilon(\mathbf{r}, \omega)$ is often called the effective permittivity. Media where the effective permittivity is a function of frequency are termed dispersive media, or frequency dispersive media.

The above concept of simple relation between flux and field can be adapted for magnetic flux and field. By a quirk of history, the magnetic flux density $\mathbf{B}$ is related to the magnetic field $\mathbf{H}$ and magnetization $\mathbf{M}$ as

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) \tag{7.1.7}$$

Defining a magnetic susceptibility χ_m such that $\mathbf{M} = \chi_m \mathbf{H}$, one gets the relationship that

$$\mathbf{B} = \mu_0(1 + \chi_m)\mathbf{H} \tag{7.1.8}$$

which is analogous to the relationship between electric flux $\mathbf{D}$ and electric field $\mathbf{E}$.

7.1.2 Anisotropic Media

For anisotropic media [34][p. 83]

$$\begin{aligned} \mathbf{D} &= \varepsilon_0 \mathbf{E} + \varepsilon_0 \overline{\boldsymbol{\chi}}_e(\omega) \cdot \mathbf{E} \\ &= \varepsilon_0 \left[\overline{\mathbf{I}} + \overline{\boldsymbol{\chi}}_e(\omega)\right] \cdot \mathbf{E} = \overline{\boldsymbol{\varepsilon}}(\omega) \cdot \mathbf{E} \end{aligned} \tag{7.1.9}$$

In the above, $\overline{\boldsymbol{\varepsilon}}$ is a 3×3 matrix also known as a tensor in electromagnetics. The above implies that $\mathbf{D}$ and $\mathbf{E}$ do not necessary point in the same direction: the meaning of anisotropy. (A tensor is a special kind of matrix that is often associated with a physical notion like the relation between two physical fields, whereas a matrix is not.)

Previously, we have assumed that χ_e to be frequency independent. This is not usually the case as all materials have χ_e's that are frequency dependent. (This will become clear later.) Also, since $\overline{\boldsymbol{\varepsilon}}(\omega)$ is frequency dependent, we should view it as a transfer function where the input is $\mathbf{E}$, and the output is $\mathbf{D}$. This implies that in the time-domain, the above relation becomes a time-convolution relation as in (7.1.4).

Similarly for conductive media, from Ohm's law,

$$\mathbf{J} = \sigma \mathbf{E}, \tag{7.1.10}$$

This can be used in Maxwell's equations in the frequency domain to yield the definition of complex permittivity. Using the above in Ampere's law in the frequency domain, we have

$$\nabla \times \mathbf{H}(\mathbf{r}) = j\omega\varepsilon \mathbf{E}(\mathbf{r}) + \sigma \mathbf{E}(\mathbf{r}) = j\omega \underset{\sim}{\varepsilon}(\omega) \mathbf{E}(\mathbf{r}) \tag{7.1.11}$$

where the complex permittivity $\underset{\sim}{\varepsilon}(\omega) = \varepsilon - j\sigma/\omega$. Notice that Ampere's law in the frequency domain with complex permittivity in (7.1.11) is no more complicated than Ampere's law for nonconductive media. The algebra for complex numbers is no more difficult than the algebra for real numbers.[1] This is one of the strengths of phasor technique.

For anisotropic conductive media, one has

$$\mathbf{J}(\omega) = \overline{\boldsymbol{\sigma}}(\omega) \cdot \mathbf{E}(\omega), \tag{7.1.12}$$

[1] Computer scientists call two systems having the same algebraic structure homomorphic. We will use the term "homomorphism" to denote such, even though its precise mathematical meaning is quite abstract. We will use the term "homomorphism" to imply similar structures.

Here, again, due to the tensorial nature of the conductivity $\overline{\boldsymbol{\sigma}}$, the electric current $\mathbf{J}$ and electric field $\mathbf{E}$ do not necessary point in the same direction.

The above assumes a local or point-wise linear time-invariant (LTI) relationship between the input and the output of a linear system. This need not be so. In fact, the most general linear time-invariant (LTI) relationship between $\mathbf{P}(\mathbf{r},t)$ and $\mathbf{E}(\mathbf{r},t)$ is

$$\mathbf{P}(\mathbf{r},t) = \varepsilon_0 \int_{-\infty}^{\infty} \iiint_{-\infty}^{\infty} \overline{\boldsymbol{\chi}}(\mathbf{r}-\mathbf{r}', t-t') \cdot \mathbf{E}(\mathbf{r}',t') d\mathbf{r}' dt' \tag{7.1.13}$$

The above is a general convolutional relationship in both space and time. The most notable fact is the non-locality of the $\mathbf{P}(\mathbf{r},t)$ that depends on the electric field $\mathbf{E}(\mathbf{r}',t')$ at $\mathbf{r}' \neq \mathbf{r}$. In the Fourier transform space, by taking Fourier transform in both space and time, the above becomes

$$\mathbf{P}(\mathbf{k},\omega) = \varepsilon_0 \overline{\boldsymbol{\chi}}(\mathbf{k},\omega) \cdot \mathbf{E}(\mathbf{k},\omega) \tag{7.1.14}$$

where

$$\overline{\boldsymbol{\chi}}(\mathbf{k},\omega) = \int_{-\infty}^{\infty} \overline{\boldsymbol{\chi}}(\mathbf{r},t) \exp(j\mathbf{k}\cdot\mathbf{r} - j\omega t) d\mathbf{r} dt \tag{7.1.15}$$

$$\mathbf{P}(\mathbf{k},\omega) = \int_{-\infty}^{\infty} \mathbf{P}(\mathbf{r},t) \exp(j\mathbf{k}\cdot\mathbf{r} - j\omega t) d\mathbf{r} dt \tag{7.1.16}$$

$$\mathbf{E}(\mathbf{k},\omega) = \int_{-\infty}^{\infty} \mathbf{E}(\mathbf{r},t) \exp(j\mathbf{k}\cdot\mathbf{r} - j\omega t) d\mathbf{r} dt \tag{7.1.17}$$

(The $d\mathbf{r}$ integral above is actually a three-fold integral with $d\mathbf{r} = dxdydz$.) Such a medium is termed spatially dispersive as well as frequency dispersive [37][p. 6], [58]. In general[2]

$$\overline{\boldsymbol{\varepsilon}}(\mathbf{k},\omega) = \varepsilon_0 (1 + \overline{\boldsymbol{\chi}}(\mathbf{k},\omega)) \tag{7.1.18}$$

where

$$\mathbf{D}(\mathbf{k},\omega) = \overline{\boldsymbol{\varepsilon}}(\mathbf{k},\omega) \cdot \mathbf{E}(\mathbf{k},\omega) \tag{7.1.19}$$

The above can be extended to magnetic field and magnetic flux yielding

$$\mathbf{B}(\mathbf{k},\omega) = \overline{\boldsymbol{\mu}}(\mathbf{k},\omega) \cdot \mathbf{H}(\mathbf{k},\omega) \tag{7.1.20}$$

for a general spatial and frequency dispersive magnetic material. In optics, most materials are non-magnetic, and hence, $\mu = \mu_0$, whereas it is quite easy to make anisotropic magnetic materials in radio and microwave frequencies, such as ferrites.

7.1.3 Bi-anisotropic Media

In the previous section, the electric flux $\mathbf{D}$ depends only on the electric field $\mathbf{E}$ and the magnetic flux $\mathbf{B}$, only on the magnetic field $\mathbf{H}$. The concept of constitutive relations can be extended to

[2] In the following, to be precise, one should replace the 1 with an identity operator, but it is generally implied.

where $\mathbf{D}$ and $\mathbf{B}$ depend on both $\mathbf{E}$ and $\mathbf{H}$. In general, one can write a general linear relationship as

$$\mathbf{D} = \overline{\boldsymbol{\varepsilon}}(\omega) \cdot \mathbf{E} + \overline{\boldsymbol{\xi}}(\omega) \cdot \mathbf{H} \tag{7.1.21}$$

$$\mathbf{B} = \overline{\boldsymbol{\zeta}}(\omega) \cdot \mathbf{E} + \overline{\boldsymbol{\mu}}(\omega) \cdot \mathbf{H} \tag{7.1.22}$$

An LTI medium where the electric flux or the magnetic flux is dependent on both $\mathbf{E}$ and $\mathbf{H}$ is known as a bi-anisotropic medium [34][p. 81].

7.1.4 Inhomogeneous Media

If any of the $\overline{\boldsymbol{\varepsilon}}$, $\overline{\boldsymbol{\xi}}$, $\overline{\boldsymbol{\zeta}}$, or $\overline{\boldsymbol{\mu}}$ is a function of position $\mathbf{r}$, the medium is termed an inhomogeneous medium or a heterogeneous medium. There are usually no simple solutions to problems associated with such media [37].

7.1.5 Uniaxial and Biaxial Media

Anisotropic optical materials are often encountered in optics. Examples of them are the biaxial and uniaxial media, and discussions of them are often found in optics books [59, 60, 61]. They are optical materials where the permittivity tensor can be written as

$$\overline{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 & 0 & 0 \\ 0 & \varepsilon_2 & 0 \\ 0 & 0 & \varepsilon_3 \end{pmatrix} \tag{7.1.23}$$

When $\varepsilon_1 \neq \varepsilon_2 \neq \varepsilon_3$, the medium is known as a biaxial medium. But when $\varepsilon_1 = \varepsilon_2 \neq \varepsilon_3$, then the medium is a uniaxial medium.

In the biaxial medium case, all three components of the electric field "feel" different permittivity constants. But in the uniaxial medium, the electric field in the xy plane feels the same permittivity constant, but the electric field in the z direction feels a different permittivity constant. As shall be shown later, different light polarization will propagate with different behaviors through such a medium. It gives rise to birefringence phenomenon which is an interesting optical phenomenon.

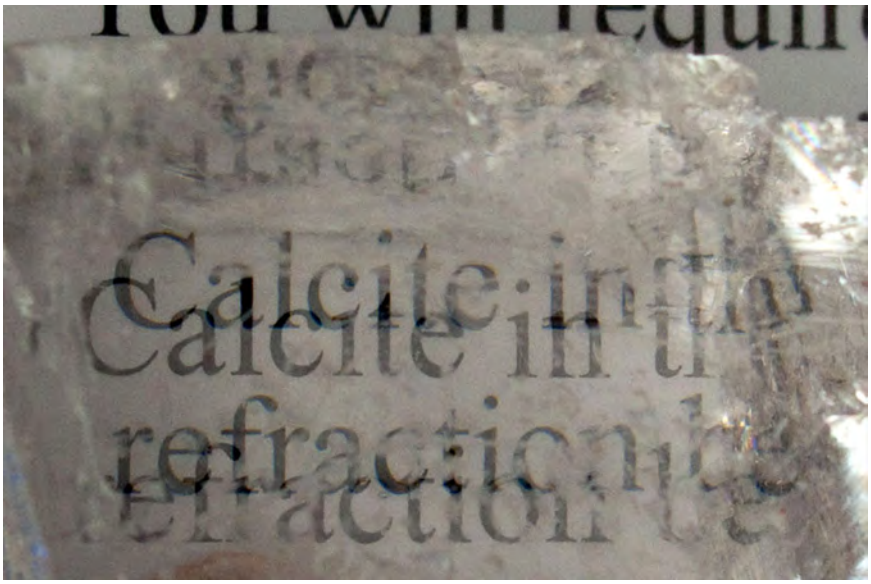


Figure 7.2: A uniaxial medium causes different light polarizations to have different velocities, and hence, refract differently.

7.1.6 Nonlinear Media

In the previous cases, we have assumed that $\overline{\chi}_e$ is independent of the field $\mathbf{E}$. The relationships between $\mathbf{P}$ and $\mathbf{E}$ can be written more generally as

$$\mathbf{P} = \varepsilon_0 \overline{\chi}_e(\mathbf{E}) \cdot \mathbf{E} \tag{7.1.24}$$

where the relationship can appear in many different forms. For nonlinear media, the relationship can be nonlinear as indicated in the above. It can be easily shown that the principle of linear superposition does not hold for the above equation, a root test of linearity. Nonlinear permittivity effect is important in optics. Here, the wavelength is short, and a small change in the permittivity or refractive index can give rise to cumulative phase delay as the wave has to propagate many wavelengths through a nonlinear optical medium [62, 63, 64]. Kerr optical nonlinearity, discovered in 1875, was one of the earliest nonlinear phenomena observed [62, 59, 34].

For magnetic materials, nonlinearity can occur in the effective permeability of the medium. In other words,

$$\mathbf{B} = \overline{\boldsymbol{\mu}}(\mathbf{H}) \cdot \mathbf{H} \tag{7.1.25}$$

This nonlinearity is important even at low frequencies, as in electric machinery designs [65, 66], and magnetic resonance imaging systems [67]. The large permeability in magnetic materials is usually due to the formation of magnetic domains which can only happen at low frequencies. The **B**-**H** relation in the metal of an electric machinery is shown in Figure 7.3. The **B**-**H** is clearly nonlinear.

The arrow indicates the time-evolution of this relationship, and clearly, it is none time-reversible. The loss of the system is related to the area of the hysterisis loop.

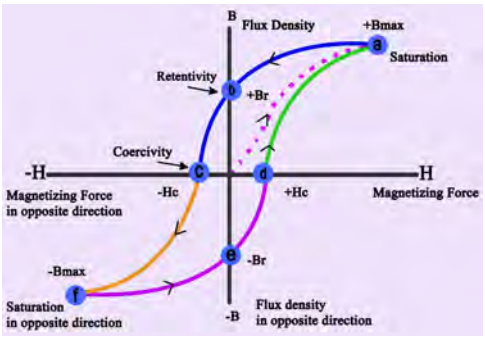


Figure 7.3: In an electric machinery, the relation between the **B** flux and the magnetic **H** is usually nonlinear, as shown in the picture. Moreover, the nonlinear system is not time-reversible, yielding a **B**-**H** relation as shown (courtesy of electricalacademia.com).

7.2 Wave Phenomenon in the Frequency Domain

We have seen the emergence of wave phenomenon in the time domain previously. Given the simplicity of the frequency domain method, it will be interesting to ask how this phenomenon presents itself for time-harmonic fields. In the frequency domain, the source-free Maxwell's equations are [34][p. 429], [68][p. 107]

$$\nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mu\mathbf{H}(\mathbf{r}) \tag{7.2.1}$$

$$\nabla \times \mathbf{H}(\mathbf{r}) = j\omega\varepsilon\mathbf{E}(\mathbf{r}) \tag{7.2.2}$$

Taking the curl of (7.2.1) and then substituting (7.2.2) into its right-hand side, one obtains

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mu\nabla \times \mathbf{H}(\mathbf{r}) = \omega^2\mu\varepsilon\mathbf{E}(\mathbf{r}) \tag{7.2.3}$$

The above is the vector Helmholtz equation. Again, using the BAC-CAB identity that

$$\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla \cdot \nabla\mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2\mathbf{E} \tag{7.2.4}$$

and that $\nabla \cdot \mathbf{E} = 0$ in a source-free medium, (7.2.3) becomes

$$(\nabla^2 + \omega^2\mu\varepsilon)\mathbf{E}(\mathbf{r}) = 0 \tag{7.2.5}$$

This is known as the Helmholtz wave equation or just the Helmholtz equation.[3]

[3]For a comprehensive review of this topic, one may read the lecture notes [47].

To see the wave phenomenon lucidly, we simply let $\mathbf{E} = \hat{x}E_x(z)$, a field pointing in the x direction, but varying only in the z direction. Evidently, $\nabla \cdot \mathbf{E}(\mathbf{r}) = \partial E_x(z)/\partial x = 0$. Then with $\partial/\partial x = 0$ and $\partial/\partial y = 0$, (7.2.5) simplifies to

$$\left(\frac{d^2}{dz^2} + k^2\right) E_x(z) = 0 \tag{7.2.6}$$

where $k^2 = \omega^2\mu\varepsilon = \omega^2/c^2$ where $c = 1/\sqrt{\mu\varepsilon}$ is the velocity of light. The general solution to (7.2.6) is of the form

$$E_x(z) = E_{0+}e^{-jkz} + E_{0-}e^{jkz} \tag{7.2.7}$$

Since time-harmonic field is assumed, one can convert the above back to the time domain using phasor technique, or by using that $E_x(z,t) = \Re e[E_x(z,\omega)e^{j\omega t}]$, yielding

$$E_x(z,t) = |E_{0+}|\cos(\omega t - kz + \alpha_+) + |E_{0-}|\cos(\omega t + kz + \alpha_-) \tag{7.2.8}$$

where we have assumed that $E_{0\pm}$ are complex numbers such that

$$E_{0\pm} = |E_{0\pm}|e^{j\alpha_\pm} \tag{7.2.9}$$

The physical picture of the above expressions can be appreciated by rewriting

$$\cos(\omega t \mp kz + \alpha_\pm) = \cos\left[\frac{\omega}{c}(ct \mp z) + \alpha_\pm\right] \tag{7.2.10}$$

where we have used the fact that $k = \frac{\omega}{c}$. The above functions are of the form $F(ct \mp x)$. As mentioned before in (3.2.14) and (3.2.15), these are traveling waves. One can see that the first term on the right-hand side of (7.2.8) is a sinusoidal plane wave traveling to the right, while the second term is a sinusoidal plane wave traveling to the left, both with velocity c.[4] The above plane wave is uniform and a constant in the xy plane but propagating in the z direction. Hence, it is also called a uniform plane wave in 1D.

Moreover, for a fixed t or at $t = 0$, the sinusoidal functions are proportional to $\cos(\mp kz + \alpha_\pm)$. This is a periodic function in z with period $2\pi/k$ which is the wavelength λ, or that

$$k = \frac{2\pi}{\lambda} = \frac{\omega}{c} = \frac{2\pi f}{c} \tag{7.2.11}$$

One can see that because c is a humongous number in free space electromagnetics, λ can be very large. You can plug in the frequency of your local AM 920 station, operating at 920 KHz, to see that λ is approximately 320 m, the size of several football fields.

The above analysis still holds true even if ϵ and μ are dispersive, but are real numbers. In this case, the velocity c of the wave is the velocity of its phase, or the phase velocity of the monochromatic, time-harmonic, or CW wave. Since ε and μ can be functions of frequency, the velocity $c = 1/\sqrt{\mu\varepsilon}$ can be different for different frequencies. As shall be shown, this gives rise to pulse distortion.

[4] We shall learn later that this is the phase velocity of the wave.

7.3 Uniform Plane Waves in 3D

By repeating the previous derivation for a homogeneous, lossless, dispersive medium, the vector Helmholtz equation for a source-free medium is given by [47]

$$\nabla \times \nabla \times \mathbf{E} - \omega^2 \mu\varepsilon \mathbf{E} = 0 \tag{7.3.1}$$

By the same derivation as before for the free-space case, since $\nabla \cdot \mathbf{E} = 0$ due to source-free medium, one has

$$\nabla^2 \mathbf{E} + \omega^2 \mu\varepsilon \mathbf{E} = 0 \tag{7.3.2}$$

The general solution to (7.3.2) is hence

$$\mathbf{E} = \mathbf{E}_0 e^{-jk_x x - jk_y y - jk_z z} = \mathbf{E}_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \tag{7.3.3}$$

where $\mathbf{k} = \hat{x}k_x + \hat{y}k_y + \hat{z}k_z$, $\mathbf{r} = \hat{x}x + \hat{y}y + \hat{z}z$ and $\mathbf{E}_0$ is a constant vector. Here, $\mathbf{k}$ is the k vector or the wave vector, while $\mathbf{r}$ is the position vector. And upon substituting (7.3.3) into (7.3.2), doing the derivatives, it is seen that

$$k_x^2 + k_y^2 + k_z^2 = \omega^2 \mu\varepsilon = \mathbf{k} \cdot \mathbf{k} \tag{7.3.4}$$

This is called the dispersion relation for a plane wave. The above is also the equation for a sphere in a 3D $\mathbf{k}$ space, which is also called the Ewald sphere.

In general, k_x, k_y, and k_z can be arbitrary and even complex numbers as long as this relation (7.3.4) is satisfied. To simplify the discussion, we will focus on the case where k_x, k_y, and k_z are all real numbers. When this is the case, the vector function in (7.3.3) represents a uniform plane wave propagating in the $\mathbf{k}$ direction. As can be seen, when $\mathbf{k} \cdot \mathbf{r}$ = constant, it is represented by all points of $\mathbf{r}$ that represents a flat plane (see Figure 7.4). This flat plane represents the constant phase wave front. By increasing the constant, we obtain different planes for progressively changing phase fronts.[5]

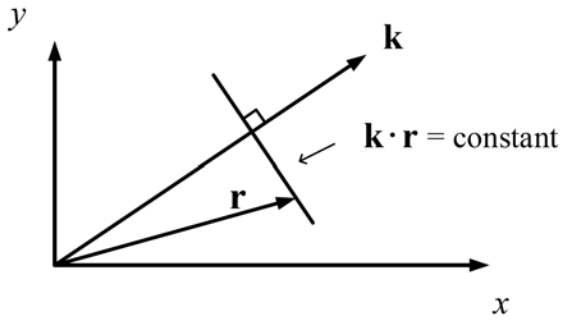


Figure 7.4: A figure showing the geometrical meaning of $\mathbf{k} \cdot \mathbf{r}$ equal to a constant. It is a flat plane that defines the wavefront of a plane wave.

[5]In the $\exp(j\omega t)$ time convention, this phase front is decreasing, whereas in the $\exp(-i\omega t)$ time convention, this phase front is increasing. The $\exp(j\omega t)$ time convention is often used in electrical engineering, while the $\exp(-i\omega t)$ time convention is used in optics and physics.

Further, since $\nabla \cdot \mathbf{E} = 0$, we can show that

$$\begin{aligned} \nabla \cdot \mathbf{E} &= \nabla \cdot \mathbf{E}_0 e^{-jk_x x - jk_y y - jk_z z} = \nabla \cdot \mathbf{E}_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \\ &= (-\hat{x}jk_x - \hat{y}jk_y - \hat{z}jk_z) \cdot \mathbf{E}_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \\ &= -j(\hat{x}k_x + \hat{y}k_y + \hat{z}k_z) \cdot \mathbf{E} = 0 \end{aligned} \tag{7.3.5}$$

or that

$$\mathbf{k} \cdot \mathbf{E}_0 = \mathbf{k} \cdot \mathbf{E} = 0 \tag{7.3.6}$$

Thus, both $\mathbf{E}$ and $\mathbf{E}_0$ are orthogonal to $\mathbf{k}$ for a uniform plane wave.

The above exercise shows that whenever $\mathbf{E}$ is a plane wave, and when the ∇ operator operates on such a vector function, one can simply let $\nabla \rightarrow -j\mathbf{k}$. Hence, in a source-free homogenous medium, Faraday's law becomes

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{7.3.7}$$

The above equation simplifies to

$$-j\mathbf{k} \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{7.3.8}$$

or that

$$\mathbf{H} = \frac{\mathbf{k} \times \mathbf{E}}{\omega\mu} \tag{7.3.9}$$

Similar to (7.3.3), we can define

$$\mathbf{H} = \mathbf{H}_0 e^{-jk_x x - jk_y y - jk_z z} = \mathbf{H}_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \tag{7.3.10}$$

Then using (7.3.3) in (7.3.9), it is clear that

$$\mathbf{H}_0 = \frac{\mathbf{k} \times \mathbf{E}_0}{\omega\mu} \tag{7.3.11}$$

We can assume that $\mathbf{E}_0$ and $\mathbf{H}_0$ are real vectors for easier visualization. Then $\mathbf{E}_0$, $\mathbf{H}_0$ and $\mathbf{k}$ form a right-handed orthogonal system, or that $\mathbf{E}_0 \times \mathbf{H}_0$ point in the direction of $\mathbf{k}$. (This also implies that $\mathbf{E}$, $\mathbf{H}$ and $\mathbf{k}$ form a right-handed orthogonal system as well.) Such a wave, where the electric field and magnetic field are transverse to the direction of propagation, is called a transverse electromagnetic (TEM) wave. Figure 7.5 shows that $\mathbf{k}\cdot\mathbf{E} = 0$, and that $\mathbf{k}\times\mathbf{E}$ points in the direction of $\mathbf{H}$ as shown in (7.3.9). Figure 7.5 also shows that $\mathbf{k}$, $\mathbf{E}$, and $\mathbf{H}$ are orthogonal to each other.

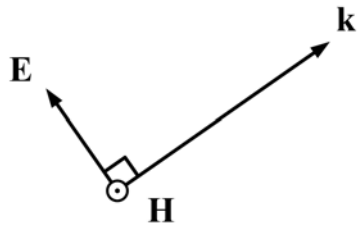


Figure 7.5: The **E**, **H**, and **k** together form a right-hand coordinate system, obeying the right-hand rule. Namely, $\mathbf{E} \times \mathbf{H}$ points in the direction of **k**.

Since in general, $\mathbf{E}_0$ and $\mathbf{H}_0$ can be complex vectors, because they are phasors, we need to show the above relationship for the more general case. From (7.3.9), one can show, using the BAC-CAB formula, assuming **k** is real, that

$$\mathbf{E} \times \mathbf{H}^* = \mathbf{E} \cdot \mathbf{E}^* \frac{\mathbf{k}}{\omega\mu} = |\mathbf{E}|^2 \frac{\mathbf{k}}{\omega\mu} \tag{7.3.12}$$

(It is important to note that the magnitude square of a complex vector is $|\mathbf{E}|^2$ is $\mathbf{E} \cdot \mathbf{E}^*$, whereas that for a real vector, it is $\mathbf{E} \cdot \mathbf{E}$. The latter definition does not guarantee positive definiteness for complex vectors.) But $\mathbf{E} \times \mathbf{H}^*$ is the direction of power flow, and it is in fact a real vector pointing in the **k** direction. This is also required by the Poynting's theorem.

Furthermore, we can show in general that[6]

$$|\mathbf{H}|^2 = \frac{|\mathbf{k} \times \mathbf{E}|^2}{(\omega\mu)^2} = \frac{\varepsilon}{\mu}|\mathbf{E}|^2 \tag{7.3.13}$$

or that

$$|\mathbf{H}| = \sqrt{\frac{\varepsilon}{\mu}}|\mathbf{E}| = \frac{1}{\eta}|\mathbf{E}| \tag{7.3.14}$$

where the quantity

$$\eta = \sqrt{\frac{\mu}{\varepsilon}} \tag{7.3.15}$$

is call the intrinsic impedance. For vacuum or free-space, it is about $377\ \Omega \approx 120\pi\ \Omega$.

Notice that the above analysis holds true as long as ε and μ are real, but they can be frequency dispersive, since we are considering a mono-chromatic or time-harmonic field. Besides, for a mono-chromatic signal, the analysis in Section 7.2 still applies except that the velocity of light is now given by $c = 1/\sqrt{\mu\varepsilon}$. As we shall see, this velocity is the phase velocity of the mono-chromatic wave. In the above, when k_x, k_y, and k_z are not all real, the wave is known as an inhomogeneous wave.[7]

[6]Since $|\mathbf{k} \times \mathbf{E}|^2 = (\mathbf{k} \times \mathbf{E}) \cdot (\mathbf{k} \times \mathbf{E}^*)$, we can use the cyclic formula $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a})$ to simplify it.

[7]The term inhomogeneous plane wave is used sometimes, but it is a misnomer since there is no more a planar wave front in this case.

Exercises for Lecture 7

Problem 7-1: For uniaxial medium, the permittivity tensor is given by:

$$\overline{\varepsilon} = \begin{pmatrix} \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon_z \end{pmatrix} \tag{E7.1}$$

Assume a plane wave propagating as

$$\mathbf{E} = \mathbf{E}_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \tag{E7.2}$$

(i) From Maxwell's equations, show that the following equation must be satisfied:

$$\mathbf{k} \times \mathbf{k} \times \mathbf{E} = -\omega^2 \mu \overline{\varepsilon} \cdot \mathbf{E} \tag{E7.3}$$

(ii) When the electric field is polarized in the xy plane, ε_z is not felt by the wave. This is called the ordinary wave. Assume that $\mathbf{k}$ is in the xz plane, show that the dispersion relation from the above equation simplifies to:

$$k_x^2 + k_z^2 = \omega^2 \mu \varepsilon \tag{E7.4}$$

(iii) When the electric field is polarized in the xz plane, ε_z is now felt by the wave. The wave is now called the extra-ordinary wave. Assume that $\mathbf{k}$ is in the xz plane, or $k_y = 0$, show that the electric flux has to be of the form:

$$\mathbf{D} = \left(\hat{x} - \hat{z}\frac{k_x}{k_z}\right) \varepsilon E_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \tag{E7.5}$$

And the corresponding electric field is:

$$\mathbf{E} = \left(\hat{x} - \hat{z}\frac{k_x \varepsilon}{k_z \varepsilon_z}\right) E_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \tag{E7.6}$$

Explain your reasoning.

(iv) From (E7.3), for the extra-ordinary wave, show that the dispersion relation can be reduced to:

$$\frac{k_x^2}{\omega^2 \mu \varepsilon_z} + \frac{k_z^2}{\omega^2 \mu \varepsilon} = 1 \tag{E7.7}$$

(v) The equations (E7.4) and (E7.7) are equations of surfaces known as k-surfaces. Please draw these two surfaces on the same graph (in 2D, it will just be a contour), and explain the physical meanings of the two surfaces.

Chapter 8

Lossy Media, Lorentz Force Law, Drude-Lorentz-Sommerfeld Model

In the previous lecture, we realized the power of phasor technique or the frequency domain analysis. The analysis of a frequency dispersive medium where ε is frequency dependent, is similar to that of free space or vacuum. The two problems are mathematically "homomorphic" to each other. In this lecture, we will generalize to the case where ε becomes a complex number, called the complex permittivity. Using phasor technique, this way of solving Maxwell's equations is still homomorphic to that of solving Maxwell's equations in free space. The analysis is greatly simplified as a result!

8.1 Plane Waves in Lossy Conductive Media

Previously, we have derived the plane wave solution for a lossless homogeneous medium. Since the algebra of complex numbers is similar to that of real numbers, the derivation can be generalized to a conductive medium by invoking mathematical "homomorphism", since the algebra of real number is similar to the algebra of complex number. In a word, in a conductive medium, one only needs to replace the permittivity with a complex permittivity, as repeated here. When conductive loss is present, $\sigma \neq 0$, and $\mathbf{J} = \sigma\mathbf{E}$. Then generalized Ampere's law becomes

$$\nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E} + \sigma\mathbf{E} = j\omega\left(\varepsilon + \frac{\sigma}{j\omega}\right)\mathbf{E} \tag{8.1.1}$$

A complex permittivity can be defined as $\underset{\sim}{\varepsilon} = \varepsilon - j\frac{\sigma}{\omega}$. Eq. (8.1.1) can be rewritten as

$$\nabla \times \mathbf{H} = j\omega\underset{\sim}{\varepsilon}\mathbf{E} \tag{8.1.2}$$

This equation is of the same form as the source-free Ampere's law in the frequency domain for a lossless medium where ε is completely real. In a conductive medium, the corresponding Helmholtz equation is

$$\left(\nabla^2 + \omega^2\mu\underset{\sim}{\varepsilon}\right)\mathbf{E} = 0 \tag{8.1.3}$$

Using the same method as before, a wave solution is[1]

$$\mathbf{E} = \mathbf{E}_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \tag{8.1.4}$$

with the dispersion relation which is now given by

$$\mathbf{k}\cdot\mathbf{k} = k_x^2 + k_y^2 + k_z^2 = \omega^2 \mu \underset{\sim}{\varepsilon} \tag{8.1.5}$$

Since $\underset{\sim}{\varepsilon}$ is complex now, k_x, k_y, and k_z cannot be all real. Equation (8.1.5) has been derived previously by assuming that $\mathbf{k}$ is a real vector. When $\mathbf{k} = \mathbf{k}' - j\mathbf{k}''$ is a complex vector, some of the previous derivations for real $\mathbf{k}$ vector may not be correct here for complex $\mathbf{k}$ vector. It is also difficult to visualize mentally a complex $\mathbf{k}$ vector that is suppose to indicate the direction with which the wave is propagating. Here, $\mathbf{k}'$ and $\mathbf{k}''$ can be vectors pointing in different directions, and the wave can decay and oscillate in different directions.

So again, for simplicity and further physical insight, we look at the simplified case where

$$\mathbf{E} = \hat{x} E_x(z) \tag{8.1.6}$$

so that $\nabla \cdot \mathbf{E} = \partial_x E_x(z) = 0$,[2] and let $\mathbf{k} = \hat{z}k = \hat{z}\omega\sqrt{\mu\underset{\sim}{\varepsilon}} = \hat{z}(k' - jk'')$. This wave is constant in the xy plane, and hence, is a plane wave. Furthermore, in this manner, we are requiring that the wave decays and propagates (or oscillates) only in the z direction. For such a simple plane wave,

$$\mathbf{E} = \hat{x}\mathbf{E}_x(z) = \hat{x} E_0 e^{-jkz} \tag{8.1.7}$$

where $k = \omega\sqrt{\mu\underset{\sim}{\varepsilon}}$ (since $\mathbf{k}\cdot\mathbf{k} = k^2 = \omega^2\mu\underset{\sim}{\varepsilon}$ is still true).

Faraday's law, by letting $\nabla \to j\mathbf{k}$, gives rise to

$$\mathbf{H} = \frac{\mathbf{k}\times\mathbf{E}}{\omega\mu} = \hat{y}\frac{kE_x(z)}{\omega\mu} = \hat{y}\sqrt{\frac{\underset{\sim}{\varepsilon}}{\mu}}E_x(z) \tag{8.1.8}$$

where the $\mathbf{k}$ vector is defined shortly after (8.1.6) above, and $k = \omega\sqrt{\mu\underset{\sim}{\varepsilon}}$, a complex number. It is seen that $\mathbf{H} = \hat{y}H_y$, and that

$$E_x/H_y = \sqrt{\frac{\mu}{\underset{\sim}{\varepsilon}}} \tag{8.1.9}$$

The above is the generalization of the intrinsic impedance defined in (7.3.15) to a lossy conductive medium.

8.1.1 High Conductivity Case

When the medium is highly conductive, $\sigma \to \infty$, and $\underset{\sim}{\varepsilon} = \varepsilon - j\frac{\sigma}{\omega} \approx -j\frac{\sigma}{\omega}$. In other words, when $|\frac{\sigma}{\omega}| \gg \varepsilon$, the conduction current dominates over the displacement current. Thus, the following approximation can be made, namely,

$$k = \omega\sqrt{\mu\underset{\sim}{\varepsilon}} \simeq \omega\sqrt{-\mu\frac{j\sigma}{\omega}} = \sqrt{-j\omega\mu\sigma} \tag{8.1.10}$$

[1] With the assumption of the wave solution below, a derivation operator $\nabla \to -j\mathbf{k}$.

[2] This condition is necessary to arrive at the Helmholtz equation (8.1.3).

Taking $\sqrt{-j} = \frac{1}{\sqrt{2}}(1-j)$, we have for a highly conductive medium that[3]

$$k \simeq (1-j)\sqrt{\frac{\omega\mu\sigma}{2}} = k' - jk'' \tag{8.1.11}$$

For a plane wave, e^{-jkz} then becomes

$$e^{-jkz} = e^{-jk'z-k''z} \tag{8.1.12}$$

By converting the above phasor back to the time domain, in the z direction, this plane wave decays exponentially as well as oscillates. The reason being that a conductive medium is lossy, and it absorbs energy from the plane wave. This is similar to resistive loss we see in the resistive circuit. The penetration depth of this wave is then

$$\delta = \frac{1}{k''} = \sqrt{\frac{2}{\omega\mu\sigma}} \tag{8.1.13}$$

This distance δ, the penetration depth, is called the skin depth of a plane wave propagating in a highly lossy conductive medium where conduction current dominates over displacement current, or that $\sigma \gg \omega\varepsilon$. This happens for radio wave propagating in the saline solution of the ocean, the good Earth, or wave propagating in highly conductive metal, like your induction cooker.

8.1.2 Low Conductivity Case

When the conductivity is low, namely, when the displacement current is larger than the conduction current, then $\frac{\sigma}{\omega\varepsilon} \ll 1$, and we have[4]

$$\begin{aligned} k &= \omega\sqrt{\mu\left(\varepsilon - j\frac{\sigma}{\omega}\right)} = \omega\sqrt{\mu\varepsilon\left(1 - \frac{j\sigma}{\omega\varepsilon}\right)} \\ &\approx \omega\sqrt{\mu\varepsilon}\left(1 - j\frac{1}{2}\frac{\sigma}{\omega\varepsilon}\right) = k' - jk'' \end{aligned} \tag{8.1.14}$$

The above is the approximation to $k = k' - jk''$ for a low conductivity medium where conduction current is much smaller than displacement current.

The term $\frac{\sigma}{\omega\varepsilon}$ is called the loss tangent of a lossy medium. It is the ratio of the conduction current to the displacement current in a lossy conductive medium. In general, in a lossy medium $\varepsilon = \varepsilon' - j\varepsilon''$, and $\varepsilon''/\varepsilon'$ is called the loss tangent of the medium. It is to be noted that in the optics and physics community, by the quirk of history, $e^{-i\omega t}$ time convention is preferred. In that case, we need to do the switch $j \to -i$, and a lossy medium is denoted by $\varepsilon = \varepsilon' + i\varepsilon''$.

[3] A function $z^{1/2}$ is known as a multi-value function. For simplicity, we assume it is a single-value function.

[4] In the following equation, we have made use of the approximation that $(1+x)^n \approx 1+nx$ when x is small, which can also be justified by Taylor series expansion.

8.2 Lorentz Force Law

The Lorentz force law is the generalization of the Coulomb's law for forces between two charges. Lorentz force law includes the presence of a magnetic field. It is given by

$$\mathbf{F} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B} \tag{8.2.1}$$

The first term on the right-hand side is the electric force similar to the statement of Coulomb's law, while the second term is the magnetic force called the $\mathbf{v} \times \mathbf{B}$ force. This law can be also written in terms of the force density $\mathbf{f}$ which is the force on the charge density, instead of force on a single charge. By so doing, we arrive at

$$\mathbf{f} = \varrho\mathbf{E} + \varrho\mathbf{v} \times \mathbf{B} = \varrho\mathbf{E} + \mathbf{J} \times \mathbf{B} \tag{8.2.2}$$

where ϱ is the charge density, and one can identified the current $\mathbf{J} = \varrho\mathbf{v}$. We shall see that the permittivity tensor of a gyrotropic medium is very much governed by this law.

Lorentz force law can also be derived from the integral form of Faraday's law, if one assumes that the law is applied to a moving loop intercepting a magnetic flux [69]. In other words, Lorentz force law and Faraday's law are commensurate with each other.

8.3 Drude-Lorentz-Sommerfeld Model

In the previous lecture, we have seen how loss can be introduced by having a conduction current flowing in a medium. Now that we have learnt the versatility of the frequency domain method and phasor technique, other loss mechanisms can be easily incorporated in the formulation.

First, let us look at the simple constitutive relation where

$$\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P} \tag{8.3.1}$$

We begin with a simple model where

$$\mathbf{P} = \varepsilon_0\chi\mathbf{E} \tag{8.3.2}$$

where χ is the electric susceptibility. To see how $\chi(\omega)$ can be derived, we will study the Drude-Lorentz-Sommerfeld model for a simplified view. This is usually just known as the Drude model or the Lorentz model in many textbooks although Sommerfeld also contributed to it. These models, the Drude, Debye, and Lorentz models, can be unified in one equation as shall be shown.

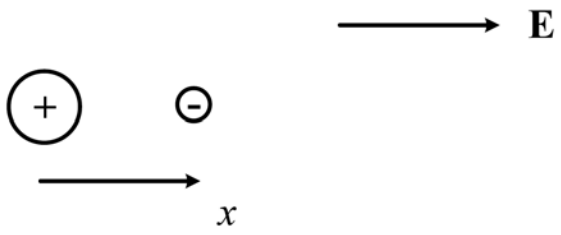


Figure 8.1: Polarization of an atom in the presence of an electric field. Here, it is assumed that the electron is weakly bound or unbound to the nucleus of the atom.

8.3.1 Cold Collisionless Plasma Medium

A plasma medium is one where the electrons of an atom have broken free from the nuclear force of the atom, so that the electrons are free to roam about. This happens in a gas that is highly heated, or in the atmosphere that is rarefied where the atoms are in an environment with very low pressure. Assuming that the free electrons are in abundance, and the atoms are far apart, we can first study, for an isolated atom, how a simple electron is driven by an electric field $\mathbf{E}$ in the absence of a magnetic field $\mathbf{B}$.[5] If the electron is free to move, then the force acting on it, from the Lorentz force law, is just $-e\mathbf{E}$ where $q = -e$ is the charge of the electron (see Figure 8.1). Using a classical model, then from Newton's law, assuming a one dimensional case, it follows that

$$m_e \frac{d^2x}{dt^2} = -eE \tag{8.3.3}$$

where the left-hand side is due to the inertial force of the mass of the electron, and the right-hand side is the electric force acting on a charge of $-e$ coulomb. Here, we assume that $\mathbf{E}$ points in the x-direction, and we neglect the vector nature of the electric field or that we assume that both x and $\mathbf{E}$ are in the same direction. Writing the above in the frequency domain for time-harmonic fields, and using phasor technique, assuming that x and E are now phasors, one gets

$$-\omega^2 m_e x = -eE \tag{8.3.4}$$

The above implies that the inertial force of the electron, given by $-\omega^2 m_e x$, is of the same polarity as the electric field force on the electron which is $-eE$. From this, we have

$$x = \frac{e}{\omega^2 m_e} E \tag{8.3.5}$$

implying that the displacement x is linearly proportional to the electric field amplitude E, or they are in phase. This, for instance, can happen in a plasma medium where the atoms are ionized, and the electrons are free to roam [71]. Hence, we assume that the positive ions are more massive, sluggish, and move very little compared to the agile electrons when an electric field is applied.

The dipole moment formed by the displaced electron away from the ion due to the electric field is then

$$p = -ex = -\frac{e^2}{\omega^2 m_e} E \tag{8.3.6}$$

for one electron. When there are N electrons per unit volume, the dipole moment density is then given by

$$P = Np = -\frac{Ne^2}{\omega^2 m_e} E \tag{8.3.7}$$

In general, $\mathbf{P}$ and $\mathbf{E}$ point in the opposite directions for this medium, and we can write

$$\mathbf{P} = -\frac{Ne^2}{\omega^2 m_e}\mathbf{E} = -\frac{{\omega_p}^2}{\omega^2}\varepsilon_0 \mathbf{E} \tag{8.3.8}$$

[5] Even if $\mathbf{B} \neq 0$, the $\mathbf{v} \times \mathbf{B}$ force is small if the velocity of the electron is much smaller than the speed of light (see Feynman [70, II-13-6]).

where we have defined ${\omega_p}^2 = Ne^2/(m_e\varepsilon_0)$ where ω_p is the plasma frequency of the medium. Then,

$$\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P} = \varepsilon_0\left(1 - \frac{{\omega_p}^2}{\omega^2}\right)\mathbf{E} \tag{8.3.9}$$

In this manner, we see that the effective permittivity of the plasma medium is

$$\varepsilon(\omega) = \varepsilon_0\left(1 - \frac{{\omega_p}^2}{\omega^2}\right) \tag{8.3.10}$$

What the above math is saying is that the electric field $\mathbf{E}$ induces a dipole moment density $\mathbf{P}$ that is negative to $\varepsilon_0\mathbf{E}$, or the vacuum part of the contribution to $\mathbf{D}$. This negative dipole density cancels the contribution to the electric flux from the vacuum $\epsilon_0\mathbf{E}$. For low frequency, the effective permittivity can be negative, disallowing the propagation of a wave as we shall see. Hence, $\varepsilon < 0$ if

$$\omega < \omega_p = \sqrt{N/(m_e\varepsilon_0)}e$$

Since $k = \omega\sqrt{\mu\varepsilon}$, if ε is negative, $k = -j\alpha$ becomes pure imaginary, and a wave such as e^{-jkz} decays exponentially as $e^{-\alpha z}$. This is also known as an evanescent wave. In other words, the wave cannot propagate through such a medium: Our ionosphere is such a medium. The ionospheric plasma shields out electromagnetic waves that are below the plasma frequency ω_p.

Therefore, it was extremely fortuitous that Marconi, in 1901, was able to send a radio signal from Cornwall, England, to Newfoundland, Canada. Nay sayers thought his experiment would not succeed as the radio signal would propagate to outer space and never to return. Fortunately so, it is the presence of the ionosphere that bounces the radio wave back to Earth, making his experiment a resounding success and a very historic one! Serendipity occurs in science and technology development more often than once: The historic experiment also heralds in the age of wireless communications.

This experiment also triggered interests into research on the ionosphere. It was an area again where Oliver Heaviside made contributions; as a result, a layer of the ionosphere is named Heaviside layer or Kennelly-Heaviside layer [72]. If you listen carefully to the Broadway musical "Cats" by Andrew Lloyd Weber, there is a mention about the Heaviside layer in one of the songs!

8.3.2 Bound Electron Case—Heuristics

Before we proceed further, we introduce a heuristic picture of how an electron would move about in a solid. A deeper understanding of this requires understanding the quantum field theory of solids [73], but an approximate picture can be obtained by studying Figure 8.2.

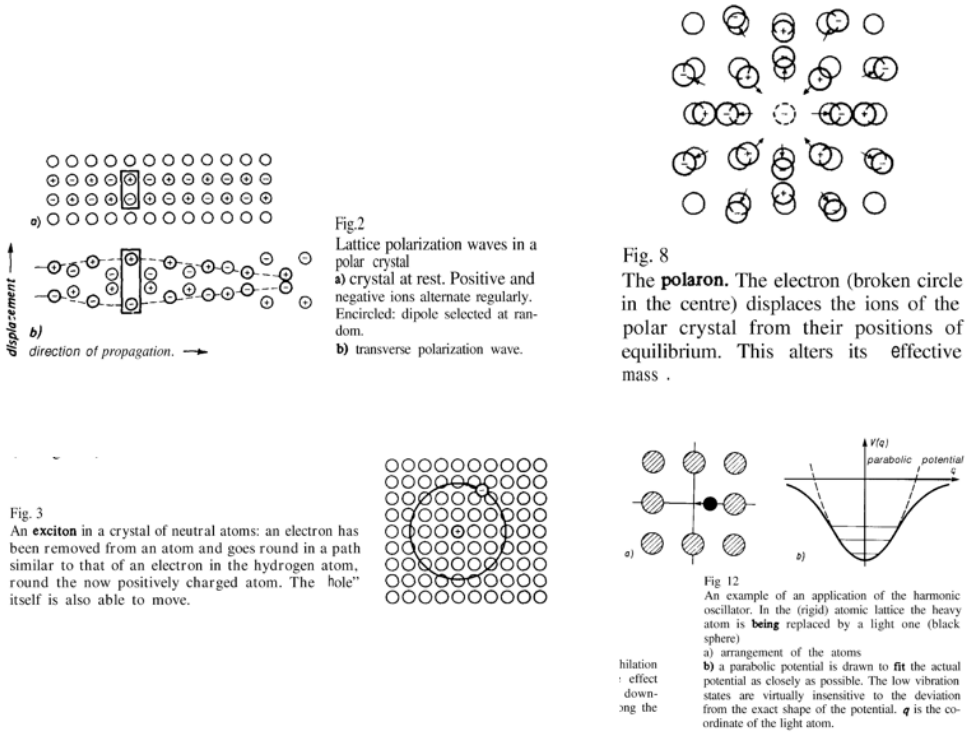


Figure 8.2: These figures are from Quantum Field Theory of Solids by H. Haken. They provide a heuristic explanation of the electromagnetic property of solids (courtesy of Haken [73]).

In Figure 8.2, the Sub-Fig. 2 illustrates the propagation of an electromagnetic wave through a polarizable medium. Sub-Fig. 3 indicates that in a semiconductor material, the electron is unbound from the nucleus forming a electron-hole pair. The electron is attracted to the hole similar to an electron around the nucleus of the hydrogen atom. Such electron-hole pair is called an exciton. Sub-Fig. 8 shows that an electron in a medium behaves like a polaron. This term is used to describe an electron with a collection of atoms, and not a free electron, because it polarizes the molecules around it. Due to its coupling to the environment, it moves about with an effective mass. Sub-Fig. 12 shows the trapping potential of an electron in the lattice. When the displacement of the electron is small from the equilibrium point, it behaves like a simple harmonic oscillator. But when its displacement is large, it becomes an anharmonic oscillator. These figures provide us with a heuristic physical understanding of the electromagnetic property of solids.

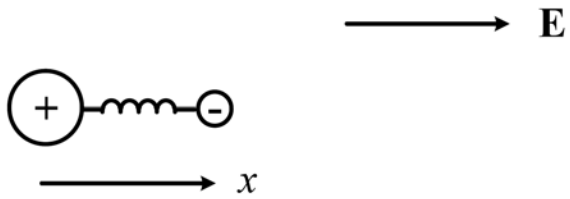


Figure 8.3: The electron with an effective mass is bound to the ion by an attractive force. This can be approximately modeled by a spring providing a restoring force to the electron.

8.3.3 Bound Electron Case—Simple Math Model

The above model of a cold collisionless plasma in (8.3.3) in Subsection 8.3.1 can be generalized to the case where the electron is bound to the ion, but the ion now provides a restoring force similar to that of a classical model or that of a spring (see Figure 8.3), namely,

$$m_e \frac{d^2x}{dt^2} + \kappa x = -eE \tag{8.3.11}$$

We assume that the ion provides a restoring force just like Hooke's law. Again, for a time-harmonic field, (8.3.11) can be solved easily using phasor technique in the frequency domain to yield

$$x = \frac{e}{(\omega^2 m_e - \kappa)} E = \frac{e}{(\omega^2 - {\omega_0}^2) m_e} E \tag{8.3.12}$$

where we have defined ${\omega_0}^2 m_e = \kappa$. The above is the typical solution of a lossless harmonic oscillator (pendulum) driven by an external force, in this case the electric field. The dipole moment due to an electric field is then

$$p = -ex = -\frac{e^2}{(\omega^2 - {\omega_0}^2) m_e} E, \qquad P = -Np = -\frac{\omega_p^2}{(\omega^2 - {\omega_0}^2)} \varepsilon E \tag{8.3.13}$$

Therefore, when the frequency is low or $\omega = 0$, polarization density P is of the same polarity as the applied electric field E, contributing to a positive dipole moment. It contributes positively to the displacement flux $\mathbf{D}$ via $\mathbf{P}$. However, when $\omega > \omega_0$, P can be out of phase with the applied field E as in the plasma medium.

8.3.4 Damping or Dissipative Case

Using a classical model, equation (8.3.11) can be generalized to the case when frictional, damping, or dissipative forces are present, or that

$$m_e \frac{d^2x}{dt^2} + m_e \Gamma \frac{dx}{dt} + \kappa x = -eE \tag{8.3.14}$$

The second term on the left-hand side is a force that is proportional to the velocity $v = dx/dt$ of the electron. This is the hall-mark of frictional force. Frictional force is due to the collision of the electrons with the background ions or lattice. Hence, it is proportional to the destruction (or change) of momentum ($m_e \frac{dx}{dt} = m_e v$) of an electron. In the average sense, the destruction of the momentum $m_e v$ is given by the product of the collision frequency Γ and the momentum. In the above, Γ has the unit of frequency, and for plasma, and conductor, it can be regarded as a collision frequency. A sanity check shows that the second term above on the left-hand side has the same unit as the first term.

Solving the above in the frequency domain, one gets

$$x = \frac{e}{(\omega^2 - j\omega\Gamma - {\omega_0}^2)m_e} E \tag{8.3.15}$$

Following the same procedure in arriving at (8.3.7), we get

$$P = \frac{\omega_p^2}{(\omega^2 - j\omega\Gamma - {\omega_0}^2)} \varepsilon E \tag{8.3.16}$$

In this, one can identify that

$$\begin{aligned} \chi(\omega) &= \frac{-Ne^2}{(\omega^2 - j\omega\Gamma - {\omega_0}^2)m_e\varepsilon_0} \\ &= -\frac{{\omega_p}^2}{\omega^2 - j\omega\Gamma - {\omega_0}^2} \end{aligned} \tag{8.3.17}$$

where ω_p is as defined before. A function with the above frequency dependence is also called a Lorentzian function. It is the hallmark of a damped harmonic oscillator.

If $\Gamma = 0$, then when $\omega = \omega_0$, one sees an infinite resonance peak exhibited by the DLS model. But in the real world, $\Gamma \neq 0$, and when Γ is small, but $\omega \approx \omega_0$, then the peak value of χ is

$$\chi \approx +\frac{{\omega_p}^2}{j\omega\Gamma} = -j\frac{{\omega_p}^2}{\omega\Gamma} \tag{8.3.18}$$

Here, χ exhibits a large negative imaginary part, the character of a dissipative medium, as in the conducting medium we have previously studied. In other words, when $\omega = \omega_0$, the DLS model is dominated by the dissipation in the medium.

8.3.5 Broad Applicability of Drude-Lorentz-Sommerfeld Model

The DLS model is a wonderful model because it can capture phenomenologically the salient feature of the physics of many electromagnetic media, even though it is purely a classical model.[6] It captures the resonance behavior of an atom absorbing energy from light excitation. When the light wave comes in at the correct frequency, it will excite electronic transition within an atom which can be approximately modeled as a resonator with behavior similar to that of a pendulum oscillator. This electronic resonances will be radiationally damped [74],[7] and the damped oscillation

[6] What we mean here is that only Newton's law has been used, and no quantum theory as yet.

[7] The oscillator radiates as it oscillates, and hence, loses energy to its environment. This causes the decay of the oscillation, just as a damped LC tank circuit losing energy to the resistor.

can be modeled by $\Gamma \neq 0$. By picking a mixture of multi-species DLS oscillators, almost any shape of absorption spectra can be curve-fitted [75] (see Figure 8.4).

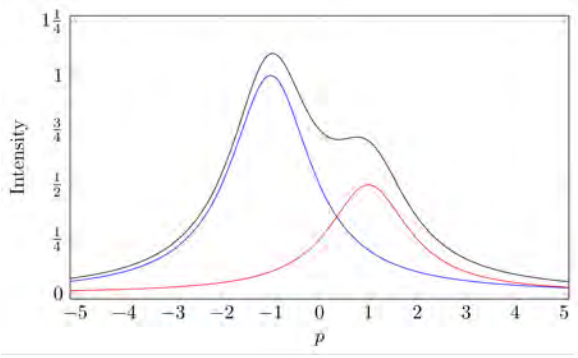


Figure 8.4: A Lorentzian has almost a bell-shape curve. By assuming multi-species of DLS oscillators in a medium, one can fit absorption spectra of almost any shape (courtesy of Wikipedia [75]).

Moreover, the above model can also be used to model molecular vibrations. In this case, the mass of the electron will be replaced by the mass of the atom involved. The damping of the molecular vibration is caused by the hindered vibration of the molecule due to interaction with other molecules [76]. The hindered rotation or vibration of water molecules when excited by microwave is the source of heat in microwave heating.

In the case of plasma, $\Gamma \neq 0$ represents the collision frequency between the free electrons and the ions, giving rise to loss. In the case of a conductor, Γ represents the collision frequency between the conduction electrons in the conduction band with the lattice of the material.[8] Also, if there is no restoring force, then $\omega_0 = 0$. This is true for the sea of electron moving in the conduction band of a medium. Besides, for sufficiently low frequency, the inertial force can be ignored. Thus, from (8.3.17), when both ω and ω_0 tend to zero, again we have[9]

$$\chi \approx -j\frac{{\omega_p}^2}{\omega\Gamma} \tag{8.3.19}$$

and

$$\varepsilon = \varepsilon_0(1+\chi) = \varepsilon_0\left(1 - j\frac{{\omega_p}^2}{\omega\Gamma}\right) \tag{8.3.20}$$

We recall that for a conductive medium, we define a complex permittivity to be

$$\varepsilon = \varepsilon_0\left(1 - j\frac{\sigma}{\omega\varepsilon_0}\right) \tag{8.3.21}$$

[8]It is to be noted that electron has a different effective mass in a crystal lattice [77, 78], and hence, the electron mass has to be changed accordingly in the DLS model.

[9]This equation is similar to (8.3.18). In both cases, collision force dominates in the equation of motion (8.3.14).

Comparing (8.3.20) and (8.3.21), we see a relation between σ, ω_p, and Γ, or that

$$\sigma = \varepsilon_0 \frac{{\omega_p}^2}{\Gamma} \tag{8.3.22}$$

The above formula for conductivity can be arrived at using collision frequency argument as is done in some textbooks such as in Streetman and Banerjee [79, p. 123].

As such, the DLS model is quite powerful: it can be used to explain a wide variety of phenomena from very low frequency to optical frequency. The electrons in many conductive materials can be modeled as a sea of free electrons moving about quite freely with an effective mass. As such, they behave like a plasma medium as shall be seen.

8.3.6 Frequency Dispersive Media—A General Discussion

The DLS model shows that, except for vacuum, all media are frequency dispersive. It is prudent to digress and discuss more on the physical meaning of a frequency dispersive medium. The relationship between electric flux and electric field, in the frequency domain, still follows the formula

$$\mathbf{D}(\omega) = \varepsilon(\omega)\mathbf{E}(\omega) \tag{8.3.23}$$

When the effective permittivity, $\varepsilon(\omega)$, is a function of frequency, it implies that the above relationship in the time domain is via convolution, viz.,

$$\mathbf{D}(t) = \varepsilon(t) \circledast \mathbf{E}(t) \tag{8.3.24}$$

Since the above represents a linear time-invariant (LTI) system [55], it implies that an input is not followed by an instantaneous output. This can be seen in equation (8.3.3) that when the electric field applies a force to the electron, the electron does not follow the field instantaneously due to its inertia in its mass. Or there is a delay between the input and the output. The reason is because an electron has a mass, and it cannot respond immediately to an applied force: or it has inertial. (In other words, the system has memory of what it was before when you try to move it.)

Even though the effective permittivity ϵ is a function of frequency, the frequency domain analysis we have done for a plane wave propagating in a dispersive medium still applies. For a monochromatic signal, it will have a velocity, called the *phase velocity*, given by $v = 1/\sqrt{\mu_0\varepsilon}$. Here, it also implies that different frequency components will propagate with different phase velocities through such a medium. Hence, a narrow pulse will spread in its width because different frequency components are not in phase after a short distance of travel. We see this in a pulse propagating in an optical fiber: the pulse loses its shape after some distance of travel.

Also, the Lorentzian function is great for data fitting, as many experimentally observed resonances have finite Q and a line width. The Lorentzian function models that well. If multiple resonances occur in a medium or an atom, then multi-species DLS model can be used. It is now clear that all media have to be frequency dispersive because of the finite mass of the electron and the inertial it has. Or there is no instantaneous response in a dielectric medium due to the finiteness of the electron mass.

Even at optical frequency, many metals, which has a sea of freely moving electrons in the conduction band, can be modeled approximately as a plasma. A metal consists of a sea of electrons

in the conduction band which are not tightly bound to the ions or the lattice. Also, in optics, the inertial force due to the finiteness of the electron mass (in this case effective mass, see Figure 8.5) can be sizeable compared to other forces. Then, $\omega_0 \ll \omega$ or that the restoring force is much smaller than the inertial force, in (8.3.17), and if Γ is small, $\chi(\omega)$ resembles that of a plasma, and ε of a metal can be negative.

Table 4.2 The effective mass m_e* of electrons in some metals.

Metal	Ag	Au	Bi	Cu	K	Li	Na	Ni	Pt	Zn
m_e^*/m_e	0.99	1.10	0.047	1.01	1.12	1.28	1.2	28	13	0.85

From *Principles of Electronic Materials and Devices, Second Edition*, S.O. Kasap (© McGraw-Hill, 2002)
http://Materials.Usask.Ca

Figure 8.5: Effective masses of electron in different metals.

8.3.7 Plasmonic Nanoparticles

When a plasmonic nanoparticle made of gold is excited by light, its response is given by (see homework assignment 8-1:)

$$\Phi_R = E_0 \frac{a^3 \cos\theta}{r^2} \frac{\varepsilon_s - \varepsilon_0}{\varepsilon_s + 2\varepsilon_0} \tag{8.3.25}$$

In a plasma, ε_s can be negative, and thus, at certain frequency, if $\varepsilon_s = -2\varepsilon_0$, then $\Phi_R \to \infty$. Gold or silver with a sea of electrons, behaves like a plasma at optical frequencies, since the inertial force in the DLS model is quite large.[10] Therefore, when light interacts with such a particle, it can sparkle brighter than normal. This reminds us of the saying "All that glitters is not gold!" even though this saying has a different intended meaning.

Ancient Romans apparently knew about the potent effect of using gold and silver nanoparticles to enhance the glistening of light in their drinking ware. These nanoparticles were impregnated in the glass or lacquer ware. By impregnating these nanoparticles in different media, the color of light will sparkle at different frequencies, and hence, the color of the glass emulsion can be changed (see website [80]).

[10] In this case, $\omega^2 \gg \omega_0^2$, and $\omega^2 \gg \omega\Gamma$; the binding force and the collision force can be ignored similar to a cold plasma.

Figure 8.6: Ancient Roman goblets whose laquer coating glisten better under lighting due to the presence of gold nanoparticles. Gold or silver at optical frequencies behaves like plasma (courtesy of Smithsonian.com).

Figure 8.7: Nanoparticles immersed in different solutions will reflect light of different colors (courtesy of nanocomposix.com).

Exercises for Lecture 8

Problem 8-1: This solution here, to a boundary value problem for Laplace equation, can be used to explain why plasmonic particles, when embedded in glass or lacquer, glitter in light. When a dielectric sphere is immersed in a static electric field as shown in the Figure 8.7, the electric field does not satisfy the boundary condition. Hence, the sphere responds by producing a dipolar potential in order to satisfy the boundary condition.

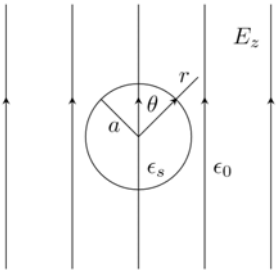


Figure 8.8: Geometry for solving this problem.

(i) Show that the potential outside the sphere can be written as

$$\Phi_{out}(\mathbf{r}) = -E_0 z + \frac{A}{r^2}\cos\theta$$

Explain the physical meanings of the first and second terms on the right-hand side of the above expression.

(ii) The potential inside the sphere can be written as

$$\Phi_{in}(\mathbf{r}) = Bz$$

where B is another unknown coefficient here. What kind of electric field corresponds to the above potential?

(iii) Now, assume that the sphere has radius a. Decide on the boundary conditions at the dielectric interface at $r = a$.

(iv) From the boundary conditions, derive the expressions for A and B and show that the second term ((i)) is of the form

$$\Phi_R = E_0 \frac{a^3 \cos\theta}{r^2} \frac{\varepsilon_s - \varepsilon_0}{\varepsilon_s + 2\varepsilon_0} \tag{8.3.1}$$

(v) Now, explain why gold plasmonic nano-particles can glitter in light.

Problem 8-2:

(i) Estimate the skin depth of the signal in your induction cooker pan. Assume that it operates around 50 KHz, and that the relative permeability μ_r is 100, and that the conductivity is about 10^7 siemens/m.

(ii) Estimate the electron density of the plasma layer in the ionosphere if it is known that radio frequency below 10 MHz cannot penetrate the ionosphere.

(iii) The conductivity of a conductive medium has been estimated to be

$$\sigma = \varepsilon_0 \frac{\omega_p^2}{\Gamma}$$

using the Drude-Lorentz-Sommerfeld model. Arrive at the same formula using collision frequency argument. **Hint**: You may find the answer in some textbooks like Streetman and Banerjee on Semiconductor Devices [79].

Chapter 9

Waves in Gyrotropic Media, Polarization

We have studied TEM uniform plane wave in Lecture 7. When the $\mathbf{k}$ vector is pointing in the z direction for instance, the electric field is polarized in the xy plane. As we shall see, assume that the electric field is polarized in the x axis, when such a wave propagates through a gyrotropic medium, its electric field rotates as it propagates. After a propagation distance, it can be polarized in other directions, such as the y direction, after emergence from this medium. Therefore, gyrotropy is an important concept in electromagnetics. In general, when a wave propagates through a gyrotropic medium, the electric field rotates changing the polarization of the wave. Our ionosphere is such a medium, and it affects radio and microwave communications between the Earth and satellites by affecting the polarization of the wave. We will study this important phenomenon in this lecture, and the general polarization of waves.

9.1 Gyrotropic Media and Faraday Rotation

A gyrotropic medium is anisotropic, and its permittivity can only be described by a tensor. Here, we will go through the beautiful mathematics, and derive the effective permittivity tensor of a gyrotropic medium, as encountered in many real-world situation such as the ionosphere, or a ferrite medium biased. These media are often biased by a static magnetic field. For instance, our ionosphere is biased by a static magnetic field arising from the Earth's magnetic field [81]. But in this derivation, in order to capture the essence of the physics with a simple model, we assume that the ionosphere has a static magnetic field polarized in the z direction, namely that $\mathbf{B} = \hat{z}B_0$. Now, the equation of motion from the Lorentz force law for an electron with $q = -e$, (in accordance with Newton's second law that $F = ma$ or force equals mass times accelration) becomes

$$m_e \frac{d\mathbf{v}}{dt} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \qquad \text{or} \qquad m_e \frac{d^2\mathbf{r}}{dt^2} = -e\left(\mathbf{E} + \frac{d\mathbf{r}}{dt} \times \mathbf{B}\right) \tag{9.1.1}$$

The first term of the force on the right-hand side is similar to Coulomb force, while the second term is usually termed the $\mathbf{v} \times \mathbf{B}$ force or the magnetic force.[1]

Next, let us assume that the electric field is polarized in the xy plane. The derivative of $\mathbf{v}$ or $d\mathbf{v}/dt$ is the acceleration of the electron, and also, $\mathbf{v} = d\mathbf{r}/dt$ where $\mathbf{r} = \hat{x}x + \hat{y}y + \hat{z}z$. Again, assuming linearity, we use phasor technique for the analysis. And in the frequency domain, the above equation in the cartesian coordinates becomes

$$m_e\omega^2 x = e(E_x + j\omega B_0 y) \tag{9.1.2}$$

$$m_e\omega^2 y = e(E_y - j\omega B_0 x) \tag{9.1.3}$$

The above constitutes two equations with two unknowns x and y. They cannot be solved easily for x and y in terms of the electric field because they correspond to a two-by-two matrix system with cross coupling between the unknowns x and y. But they can be simplified and decoupled as follows: We can multiply (9.1.3) by $\pm j$ and add it to (9.1.2) to get two decoupled equations [82]:

$$m_e\omega^2(x + jy) = e[(E_x + jE_y) + \omega B_0(x + jy)] \tag{9.1.4}$$

$$m_e\omega^2(x - jy) = e[(E_x - jE_y) - \omega B_0(x - jy)] \tag{9.1.5}$$

In the above, if we take the new unknowns to be $x \pm jy$, the two equations are decoupled with respect to these two unknowns. Now, defining new variables such that

$$s_\pm = x \pm jy \tag{9.1.6}$$

$$E_\pm = E_x \pm jE_y \tag{9.1.7}$$

then (9.1.4) and (9.1.5) become

$$m_e\omega^2 s_\pm = e(E_\pm \pm \omega B_0 s_\pm) \tag{9.1.8}$$

Thus, solving the above yields

$$s_\pm = \frac{e}{m_e\omega^2 \mp eB_0\omega} E_\pm = C_\pm E_\pm \tag{9.1.9}$$

where

$$C_\pm = \frac{e}{m_e\omega^2 \mp eB_0\omega} \tag{9.1.10}$$

[2]

Next, one can define $P_x = -Nex$, $P_y = -Ney$, and that $P_\pm = P_x \pm jP_y = -Nes_\pm$. Then it can be shown that

$$P_\pm = \varepsilon_0\chi_\pm E_\pm \tag{9.1.11}$$

[1]For a plane wave, it can be shown that the $\mathbf{v} \times \mathbf{B}$ force is of order v/c smaller than the Coulomb force, which is termed relativistically small [70, II-13-6]. But it can be made artificially large by a permanent magnet.

[2]By this manipulation, the above equations (9.1.2) and (9.1.3) transform to new equations where there is no cross coupling between $s_\pm$ and $E_\pm$. The mathematical parlance for this is the diagonalization of a matrix equation [83]. Thus, the new equation can be solved easily.

which is a scalar equation! The expression for $\chi_\pm$ can be derived, and they are given as

$$\chi_\pm = -\frac{NeC_\pm}{\varepsilon_0} = -\frac{Ne}{\varepsilon_0}\frac{e}{m_e\omega^2 \mp eB_o\omega} = -\frac{{\omega_p}^2}{\omega^2 \mp \Omega\omega} \tag{9.1.12}$$

where Ω and ω_p are the cyclotron frequency[3] and plasma frequency, respectively, viz.,

$$\Omega = \frac{eB_0}{m_e}, \quad {\omega_p}^2 = \frac{Ne^2}{m_e\varepsilon_0} \tag{9.1.13}$$

Notice that at the cyclotron frequency, $|\chi_\pm| \to \infty$, a hallmark of resonance behavior. In other words, $P_\pm$ is finite even when $E_\pm = 0$, or a solution exists to the equation of motion (9.1.1) without a forcing term, which in this case is the electric field. Thus, at this frequency, the solution blows up if the forcing term, $E_\pm$ is not zero. This is like what happens at the resonance of an LC tank circuit whose current or voltage tends to infinity when the forcing term, like the voltage or current, is nonzero. In a word, the solution exists even when the forcing term is zero, similar to the homogeneous solution of an ordinary differential equation.

In order to derive the permittivity tensor in the cartesian coordinates, we first need to express the original variables $P_x,\ P_y,\ E_x,\ E_y$ in terms of $P_\pm$ and $E_\pm$. With the help of (9.1.11), we arrive at

$$\begin{aligned} P_x = \frac{P_+ + P_-}{2} = \frac{\varepsilon_0}{2}(\chi_+E_+ + \chi_-E_-) &= \frac{\varepsilon_0}{2}[\chi_+(E_x + jE_y) + \chi_-(E_x - jE_y)] \\ &= \frac{\varepsilon_0}{2}[(\chi_+ + \chi_-)E_x + j(\chi_+ - \chi_-)E_y] \end{aligned} \tag{9.1.14}$$

$$\begin{aligned} P_y = \frac{P_+ - P_-}{2j} = \frac{\varepsilon_0}{2j}(\chi_+E_+ - \chi_-E_-) &= \frac{\varepsilon_0}{2j}[\chi_+(E_x + jE_y) - \chi_-(E_x - jE_y)] \\ &= \frac{\varepsilon_0}{2j}[(\chi_+ - \chi_-)E_x + j(\chi_+ + \chi_-)E_y] \end{aligned} \tag{9.1.15}$$

The above relationship in cartesian coordinates can be expressed using a tensor where

$$\mathbf{P} = \varepsilon_0\overline{\boldsymbol{\chi}} \cdot \mathbf{E} \tag{9.1.16}$$

where $\mathbf{P} = [P_x, P_y]$, and $\mathbf{E} = [E_x, E_y]$. From (9.1.14) and (9.1.15) above, $\overline{\boldsymbol{\chi}}$ is a tensor of the form

$$\overline{\boldsymbol{\chi}} = \frac{1}{2}\begin{pmatrix} (\chi_+ + \chi_-) & j(\chi_+ - \chi_-) \\ -j(\chi_+ - \chi_-) & (\chi_+ + \chi_-) \end{pmatrix} = \begin{pmatrix} -\frac{{\omega_p}^2}{\omega^2 - \Omega^2} & -j\frac{{\omega_p}^2\Omega}{\omega(\omega^2 - \Omega^2)} \\ j\frac{{\omega_p}^2\Omega}{\omega(\omega^2 - \Omega^2)} & -\frac{{\omega_p}^2}{\omega^2 - \Omega^2} \end{pmatrix} \tag{9.1.17}$$

Notice that in the above, when the $\mathbf{B}$ field is turned off or $\Omega = 0$, then $\overline{\boldsymbol{\chi}}$ above is diagonalize, and it resembles an isotropic medium of a collisionless, cold plasma again.

Consequently, for the $\mathbf{B} \neq 0$ case, the above can be generalized to 3D to give

$$\overline{\boldsymbol{\chi}} = \begin{bmatrix} \chi_0 & j\chi_1 & 0 \\ -j\chi_1 & \chi_0 & 0 \\ 0 & 0 & \chi_p \end{bmatrix} \tag{9.1.18}$$

[3]This is also called the gyrofrequency.

where $\chi_p = -\omega_p^2/\omega^2$. Notice that since we assume that $\mathbf{B} = \hat{z}B_0$, the z component of (9.1.1) is unaffected by the $\mathbf{v} \times \mathbf{B}$ force. Hence, the electron moving in the z direction is similar to that in a cold collisionless plasma.

Using the fact that $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P} = \varepsilon_0(\overline{\mathbf{I}} + \overline{\boldsymbol{\chi}}) \cdot \mathbf{E} = \overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}$, the above implies that

$$\overline{\boldsymbol{\varepsilon}} = \varepsilon_0 \begin{bmatrix} 1+\chi_0 & j\chi_1 & 0 \\ -j\chi_1 & 1+\chi_0 & 0 \\ 0 & 0 & 1+\chi_p \end{bmatrix} \tag{9.1.19}$$

Now, $\overline{\boldsymbol{\varepsilon}}$ is that of an anisotropic medium, of which a gyrotropic medium belongs. Please notice that the above tensor is a hermitian tensor. We shall learn later that this is the property of a lossless medium.

Another characteristic of a gyrotropic medium is that a linearly polarized wave will rotate when passing through it. This is the Faraday rotation effect [82], which we shall learn more later. This phenomenon poses a severe problem for Earth-to-satellite communications, using linearly polarized wave as it requires the alignment of the Earth-to-satellite antennas. Thank goodness, this can be avoided using a rotatingly polarized wave, called a circularly polarized wave that we shall learn in the next section.

As we have learnt, the ionosphere affects our communication systems two ways: It acts as a mirror for low-frequency electromagnetic or radio waves (making the experiment of Marconi a rousing success). It also affects the polarization of the wave. But the ionosphere of the Earth and the density of electrons that are ionized is highly dependent on temperature, and the effect of the Sun. The fluctuation of particles in the ionosphere gives rise to scintillation effects due to electron motion and collision that affect radio wave communication systems [84].

9.2 Wave Polarization

Studying wave polarization is very important for communication purposes [34]. A wave whose electric field is pointing in the x direction while propagating in the z direction is called a linearly polarized (LP) wave. The same can be said of one with electric field polarized in the y direction. It turns out that a linearly polarized wave experiences Faraday rotation when it propagates through the ionosphere. For instance, an x polarized wave can become a y polarized wave due to Faraday rotation. So its polarization becomes ambiguous as the wave propagates through the ionosphere: to overcome this, Earth to satellite communication is done with circularly polarized (CP) waves [85]. So even if the electric field vector is rotated by Faraday's rotation, it remains to be a CP wave. We will study these polarized waves next. Later, we will study how to make antennas that radiate linearly or circularly polarized waves.

9.2.1 General Polarizations—Elliptical and Circular Polarizations

We can write a general uniform plane wave propagating in the z direction in the time domain for simplicity as

$$\mathbf{E} = \hat{x}E_x(z,t) + \hat{y}E_y(z,t) \tag{9.2.1}$$

Clearly, $\nabla \cdot \mathbf{E} = 0$, and $E_x(z,t)$ and $E_y(z,t)$, by the principle of linear superposition, are solutions to the one-dimensional wave equation. For a time harmonic field, the two components may not be in phase, and in general, we have for time domain that

$$E_x(z,t) = E_1 \cos(\omega t - \beta z) \tag{9.2.2}$$
$$E_y(z,t) = E_2 \cos(\omega t - \beta z + \alpha) \tag{9.2.3}$$

where α denotes the phase difference between these two wave components. We shall study how the linear superposition of these two components behaves for different α's. To start, we observe this field at $z = 0$. Then

$$\mathbf{E} = \hat{x} E_1 \cos(\omega t) + \hat{y} E_2 \cos(\omega t + \alpha) \tag{9.2.4}$$

For $\alpha = \frac{\pi}{2}$

$$E_x = E_1 \cos(\omega t),\ E_y = E_2 \cos(\omega t + \pi/2) \tag{9.2.5}$$

Next, we evaluate the above for different ωt's

$$\omega t = 0, \qquad E_x = E_1, \qquad E_y = 0 \tag{9.2.6}$$
$$\omega t = \pi/4, \qquad E_x = E_1/\sqrt{2}, \qquad E_y = -E_2/\sqrt{2} \tag{9.2.7}$$
$$\omega t = \pi/2, \qquad E_x = 0, \qquad E_y = -E_2 \tag{9.2.8}$$
$$\omega t = 3\pi/4, \qquad E_x = -E_1/\sqrt{2}, \qquad E_y = -E_2/\sqrt{2} \tag{9.2.9}$$
$$\omega t = \pi, \qquad E_x = -E_1, \qquad E_y = 0 \tag{9.2.10}$$

The tip of the vector field $\mathbf{E}$ traces out an ellipse as show in Figure 9.1. With the left-hand thumb pointing in the z direction, the direction of propagation, and the wave rotating in the direction of the fingers, such a wave is called left-hand elliptically polarized (LHEP) wave.

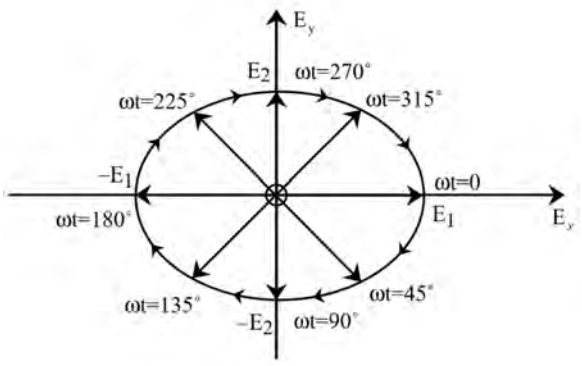


Figure 9.1: If one follows the tip of the electric field vector, it traces out an ellipse as a function of time t.

When $E_1 = E_2$, the ellipse becomes a circle, and we have a left-hand circularly polarized (LHCP) wave. When $\alpha = -\pi/2$, the wave rotates in the counter-clockwise direction, and the wave is either right-hand elliptically polarized (RHEP), or right-hand circularly polarized (RHCP) wave depending on the ratio of E_1/E_2. Figure 9.2 shows the different polarizations of the wave for different phase differences and amplitude ratio. Figure 9.3 shows a graphic picture of a CP wave propagating through space.

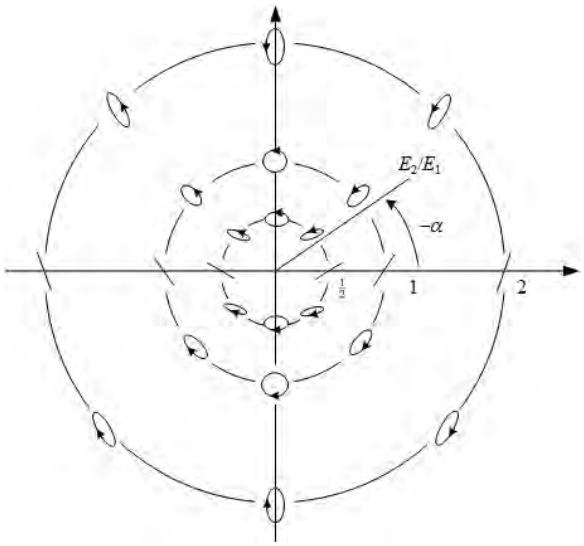


Figure 9.2: Due to different phase difference between the E_x and E_y components of the field, and their relative amplitudes E_2/E_1, different polarizations will ensue. The arrow indicates the direction of rotation of the field vector.

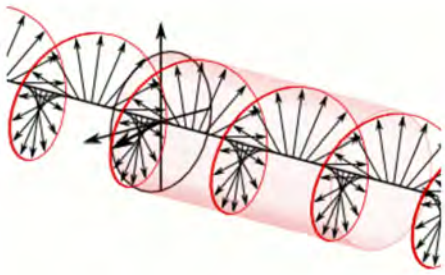


Figure 9.3: The rotation of the field vector of a right-hand circular polarization wave as it propagates in the right direction [86] (courtesy of Wikipedia).

9.2.2 Arbitrary Polarization Case and Axial Ratio[4]

As seen before, the tip of the field vector traces out an ellipse in space as it propagates. The axial ratio (AR) is the ratio of the major axis to the minor axis of this ellipse. It is an important figure of merit for designing CP (circularly polarized) antennas (antennas that will radiate circularly polarized waves). The closer is this ratio to 1, the better is the antenna design. We will discuss the general polarization and the axial ratio of a wave.

For the general case for arbitrary α, we let

$$E_x = E_1 \cos\omega t, \qquad E_y = E_2 \cos(\omega t + \alpha) = E_2(\cos\omega t \cos\alpha - \sin\omega t \sin\alpha) \tag{9.2.11}$$

Then from the above, expressing E_y in terms of E_x, one gets

$$E_y = \frac{E_2}{E_1} E_x \cos\alpha - E_2 \left[1 - \left(\frac{E_x}{E_1}\right)^2\right]^{1/2} \sin\alpha \tag{9.2.12}$$

Rearranging and squaring, we get

$$a{E_x}^2 - bE_xE_y + c{E_y}^2 = 1 \tag{9.2.13}$$

where

$$a = \frac{1}{{E_1}^2 \sin^2\alpha}, \quad b = \frac{2\cos\alpha}{E_1E_2 \sin^2\alpha}, \quad c = \frac{1}{{E_2}^2 \sin^2\alpha} \tag{9.2.14}$$

After letting $E_x \to x$, and $E_y \to y$, equation (9.2.13) is of the form,

$$ax^2 - bxy + cy^2 = 1 \tag{9.2.15}$$

which is the equation of an ellipse. The equation of an ellipse in its self coordinates is

$$\left(\frac{x'}{A}\right)^2 + \left(\frac{y'}{B}\right)^2 = 1 \tag{9.2.16}$$

where A and B are axes of the ellipse as shown in Figure 9.4. We can transform the above back to the (x, y) coordinates to get (9.2.15). To this end, we let

$$x' = x\cos\theta - y\sin\theta \tag{9.2.17}$$

$$y' = x\sin\theta + y\cos\theta \tag{9.2.18}$$

to get

$$x^2\left(\frac{\cos^2\theta}{A^2} + \frac{\sin^2\theta}{B^2}\right) - xy\sin 2\theta\left(\frac{1}{A^2} - \frac{1}{B^2}\right) + y^2\left(\frac{\sin^2\theta}{A^2} + \frac{\cos^2\theta}{B^2}\right) = 1 \tag{9.2.19}$$

[4]This section is mathematically complicated. It can be skipped on first reading.

Comparing (9.2.13) and (9.2.19), one gets

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\cos\alpha E_1 E_2}{{E_2}^2 - {E_1}^2}\right) \tag{9.2.20}$$

$$\text{AR} = \left(\frac{1+\Delta}{1-\Delta}\right)^{1/2} > 1 \tag{9.2.21}$$

where AR is the axial ratio and

$$\Delta = \left(1 - \frac{4{E_1}^2{E_2}^2\sin^2\alpha}{\left({E_1}^2 + {E_2}^2\right)^2}\right)^{1/2} \tag{9.2.22}$$

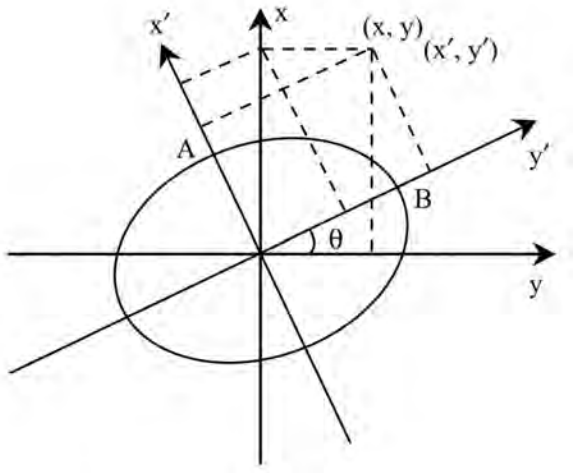


Figure 9.4: This figure shows the parameters used to derive the axial ratio (AR) of an elliptically polarized wave.

9.3 Polarization and Power Flow

For a linearly polarized wave in the time domain,[5]

$$\mathbf{E} = \hat{x}E_0\cos(\omega t - \beta z), \quad \mathbf{H} = \hat{y}\frac{E_0}{\eta}\cos(\omega t - \beta z) \tag{9.3.1}$$

Hence, the instantaneous power we have learnt previously in Section 5.3 becomes

$$\mathbf{S}(t) = \mathbf{E}(t)\times\mathbf{H}(t) = \hat{z}\frac{{E_0}^2}{\eta}\cos^2(\omega t - \beta z) \tag{9.3.2}$$

[5]In the course, β and k are used interchangeably. The former is preferred in microwave and the latter in optics.

indicating that for a linearly polarized wave, the instantaneous power is function of both time and space. It travels as lumps of energy through space. In the above E_0 is the amplitude of the linearly polarized wave. Moreover, taking the time average of the above, we have

$$\langle \mathbf{S}(t) \rangle = \hat{z} \frac{{E_0}^2}{2\eta} \tag{9.3.3}$$

Next, we look at power flow for for elliptically and circularly polarized waves. It is to be noted that in the phasor world or frequency domain, (9.2.1) now becomes

$$\mathbf{E}(z,\omega) = \hat{x} E_1 e^{-j\beta z} + \hat{y} E_2 e^{-j\beta z + j\alpha} \tag{9.3.4}$$

For LHEP wave, $E_1 \neq E_2$ and $\alpha = \pi/2$,

$$\mathbf{E}(z,\omega) = e^{-j\beta z}(\hat{x} E_1 + j\hat{y} E_2) \tag{9.3.5}$$

whereas for LHCP wave, $E_1 = E_2$ and $\alpha = \pi/2$

$$\mathbf{E}(z,\omega) = e^{-j\beta z} E_1(\hat{x} + j\hat{y}) \tag{9.3.6}$$

For RHEP wave, the above becomes

$$\mathbf{E}(z,\omega) = e^{-j\beta z}(\hat{x} E_1 - j\hat{y} E_2) \tag{9.3.7}$$

whereas for RHCP wave, it is

$$\mathbf{E}(z,\omega) = e^{-j\beta z} E_1(\hat{x} - j\hat{y}) \tag{9.3.8}$$

Focussing on the circularly polarized wave,

$$\mathbf{E} = (\hat{x} \pm j\hat{y}) E_0 e^{-j\beta z} \tag{9.3.9}$$

Using that $\boldsymbol{\beta} = \hat{z}\beta$, and letting $\nabla \to -j\boldsymbol{\beta}$, Faraday's law becomes

$$\mathbf{H} = \frac{\boldsymbol{\beta} \times \mathbf{E}}{\omega\mu} \tag{9.3.10}$$

And then

$$\mathbf{H} = (\mp\hat{x} - j\hat{y}) j \frac{E_0}{\eta} e^{-j\beta z} \tag{9.3.11}$$

where $\eta = \sqrt{\mu/\varepsilon}$ is the intrinsic impedance of the medium. Therefore,

$$\mathbf{E}(z,t) = \hat{x} E_0 \cos(\omega t - \beta z) \pm \hat{y} E_0 \sin(\omega t - \beta z) \tag{9.3.12}$$

$$\mathbf{H}(z,t) = \mp\hat{x} \frac{E_0}{\eta} \sin(\omega t - \beta z) + \hat{y} \frac{E_0}{\eta} \cos(\omega t - \beta z) \tag{9.3.13}$$

Then the instantaneous power becomes

$$\mathbf{S}(z,t) = \mathbf{E}(z,t) \times \mathbf{H}(z,t) = \hat{z} \frac{{E_0}^2}{\eta} \cos^2(\omega t - \beta z) + \hat{z} \frac{{E_0}^2}{\eta} \sin^2(\omega t - \beta z) = \hat{z} \frac{{E_0}^2}{\eta} \tag{9.3.14}$$

In other words, a CP wave delivers constant instantaneous power independent of space and time, as opposed to a linearly polarized wave which delivers a non-constant instantaneous power as shown in (9.3.2). Moreover, taking the time average of the above, we have

$$\langle \mathbf{S}(z,t) \rangle = \hat{z} \frac{{E_0}^2}{\eta} \tag{9.3.15}$$

which is independent of z, t.

It is to be noted that the complex Poynting's vector for a lossless medium

$$\underset{\sim}{\mathbf{S}} = \mathbf{E} \times \mathbf{H}^* \tag{9.3.16}$$

is real and constant independent of space both for linearly, circularly, and elliptically polarized waves. That is if we were to go through the exercise to obtain $\underset{\sim}{\mathbf{S}}$ for the general case, we will let

$$\mathbf{E} = (\hat{x} E_1 \pm j \hat{y} E_2) e^{-j\beta z} \tag{9.3.17}$$

The corresponding magnetic field can be found as

$$\mathbf{H} = \frac{\boldsymbol{\beta} \times \mathbf{E}}{\omega \mu} = \frac{\beta}{\omega \mu} (\hat{y} E_1 \mp j \hat{x} E_2) e^{-j\beta z} \tag{9.3.18}$$

Using the above, we find that the complex Poynting's vector as

$$\underset{\sim}{\mathbf{S}} = \mathbf{E} \times \mathbf{H}^* = \frac{\beta}{\omega \mu} \hat{z} (|E_1|^2 + |E_2|^2) \tag{9.3.19}$$

Then the time-average power density is

$$\langle \mathbf{S} \rangle = \frac{1}{2} \Re e \left\{ \underset{\sim}{\mathbf{S}} \right\} = \frac{1}{2\eta} \hat{z} (|E_1|^2 + |E_2|^2) \tag{9.3.20}$$

When $E_1 = E_2 = E_0$, the above becomes

$$\langle \mathbf{S} \rangle = \frac{1}{2} \Re e \left\{ \underset{\sim}{\mathbf{S}} \right\} = \frac{1}{\eta} \hat{z} |E_0|^2 \tag{9.3.21}$$

which is the same as in (9.3.14).

When $E_2 = 0$ for a linearly polarized wave, and $E_1 = E_0$, we have

$$\langle \mathbf{S} \rangle = \frac{1}{2} \Re e \left\{ \underset{\sim}{\mathbf{S}} \right\} = \frac{1}{2\eta} \hat{z} |E_0|^2 \tag{9.3.22}$$

This is the same as what we have found before in (9.3.3). Notice that the Poynting's vector is a constant independent of z. This is because there is no reactive power in a plane wave of any polarization: the stored energy in the plane wave cannot be returned to the source!

Exercises for Lecture 9

Problem 9-1: Show that (9.1.17) becomes that of collisionless cold plasma when the **B** field is turned off.

Problem 9-2: This problem will show the Faraday rotation of a linearly polarized wave as it propagates through the ionosphere. This happens because the ionosphere is a gyrotropic medium. Its permittivity tensor, as shown in this lecture, can be expressed as:

$$\overline{\varepsilon} = \begin{bmatrix} \varepsilon_x & j\varepsilon_g & 0 \\ -j\varepsilon_g & \varepsilon_y & 0 \\ 0 & 0 & \varepsilon_z \end{bmatrix} \tag{E9.1}$$

(i) From the lecture notes, surmise the forms of $\varepsilon_x, \varepsilon_y, \varepsilon_z, \varepsilon_g$. Assume all these quantities are real, how would you classify this matrix that corresponds to this medium. You will learn later that such a matrix corresponds to a lossless medium.

(ii) Assume that the wave is propagating in the z direction, given as

$$\mathbf{E} = \hat{x} E_0 e^{-jkz}$$

Show that this is the solution to Maxwell's equations as well as the vector Helmholtz equation.

(iii) Show that a linearly polarized wave can be split into a sum of a right-hand circularly polarized (RHCP) wave and a left-hand circularly polarized (LHCP) wave. In other words, show that

$$\mathbf{E} = \hat{x} E_0 e^{-jkz} = \mathbf{E}_{LHCP} + \mathbf{E}_{RHCP}$$

And find the forms of $\mathbf{E}_{LHCP}$ and $\mathbf{E}_{RHCP}$.

(iv) Show that these two waves RHCP and LHCP will propagate with different wave numbers k_+ and k_- or phase velocity in a gyrotropic medium. Find these wave numbers.

(v) Show that the wave rotates (known as Faraday rotation) as it propagates in the z direction as shown above. After a distance d in the gyrotropic medium, it becomes rotated and the wave acquires a y component as well. Therefore, the field vector is tilted with respect to the x axis.

(vi) Show that the tilt angle, or the angle of Faraday rotation is:

$$\theta_F = \frac{k_+ - k_-}{2} d$$

Chapter 10

Momentum, Complex Poynting's Theorem, Lossless Condition, Energy Density

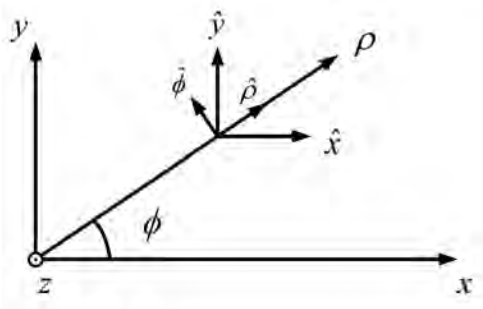


Figure 10.1: The local coordinates used to describe a circularly polarized wave: In cartesian and polar coordinates.

In the last lecture, we study circularly polarized waves as well as linearly polarized waves. In addition, these waves can carry power giving rise to power flow. But in addition to carrying power, a travelling wave also carries momentum: for a linearly polarized wave, it carries linear momentum in the direction of the propagation of the traveling wave. But for a circularly polarized wave, it carries angular momentum as well.[1]

[1] At this juncture, you will have to recalibrate your electrical engineering training and learn to think that energy flow and momentum are not carried by electron flow except for some special cases. Most often than not, they are carried by photons or the electromagnetic fields.

Previously, we have studied complex power and the complex Poynting's theorem in the frequency domain with phasors. Here, we will derive the lossless conditions for the permittivity and permeability tensors under which a medium is lossless. As we have shown in the instantaneous Poynting's theorem, energy density is well defined for a lossless dispersionless medium, but we will assumes a different formula when the medium is dispersive. From the Drude-Lorentz-Sommerfeld model, it is seen that a medium is never dispersionless save for vacuum. In other words, it $\mathbf{D}(\omega) = \varepsilon(\omega)\mathbf{E}(\omega)$, if ε is a function of frequency, the change in $\mathbf{D}$ cannot be instantaneous if there is a sudden change in $\mathbf{E}$.

10.1 Spin Angular Momentum and Cylindrical Vector Beam

In this section, we will study the spin angular momentum of a circularly polarized (CP) wave. It is to be noted that in cylindrical coordinates, as shown in Figure 10.1, $\hat{x} = \hat{\rho}\cos\phi - \hat{\phi}\sin\phi$, $\hat{y} = \hat{\rho}\sin\phi + \hat{\phi}\cos\phi$, then a CP field is proportional to

$$(\hat{x} \pm j\hat{y}) = \hat{\rho}e^{\pm j\phi} \pm j\hat{\phi}e^{\pm j\phi} = e^{\pm j\phi}(\hat{\rho} \pm \hat{\phi}) \tag{10.1.1}$$

Therefore, with the $e^{\pm j\phi}$ dependence, the $\hat{\rho}$ and $\hat{\phi}$ of a CP wave is also an azimuthal traveling wave in the $\hat{\phi}$ direction in addition to being a traveling wave $e^{-j\beta z}$ in the $\hat{z}$ direction. This is obviated by rewriting

$$e^{-j\phi} = e^{-jk_\phi \rho\phi} \tag{10.1.2}$$

where $k_\phi = 1/\rho$ is the azimuthal wave number, and $\rho\phi$ is the arc length traversed by the azimuthal wave. Notice that the wavenumber k_ϕ is dependent on ρ: the larger the ρ, the smaller is k_ϕ, and hence, the larger the azimuthal wavelength. Thus, the wave possesses angular momentum called the spin angular momentum (SAM), just as a traveling wave, $e^{-j\beta z}$ possesses linear angular momentum in the $\hat{z}$ direction.

In optics research, the generation of cylindrical vector beam is in vogue. Figure 10.2 shows a method to generate such a beam. A CP light passes through a radial analyzer that will only allow the radial component of (10.1.1) to be transmitted. Then a spiral phase element (SPE) compensates for the $\exp(\pm j\phi)$ phase shift in the azimuthal direction. Finally, the light is a cylindrical vector beam which is radially polarized without spin angular momentum. Such a beam has been found to have nice focussing property, and hence, has aroused researchers' interest in the optics community [87].

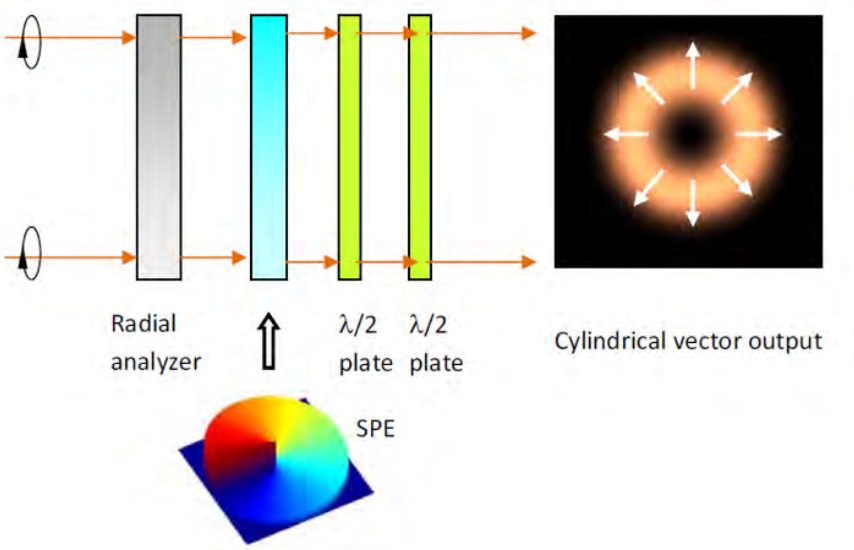


Figure 10.2: A cylindrical vector beam can be generated experimentally. The spiral phase element (SPE) compensates for the $\exp(\pm j\phi)$ phase shift (courtesy of Zhan, Q. [87]). The half-wave plate rotates the polarization of a wave by 90 degrees.

10.2 Momentum Density of Electromagnetic Field

We have seen that a traveling wave carries power and has energy density associated with it. In other words, the moving or traveling energy density gives rise to power flow. It turns out that a traveling wave also carries a momentum with it. The momentum density of electromagnetic field is given by

$$\mathbf{G} = \mathbf{D} \times \mathbf{B} \tag{10.2.1}$$

also called the momentun density vector. With it, one can derive momentum conservation theorem [34, p. 59] [49]. The derivation is rather long, but we will justify the above formula and simplify the derivation using the particle or corpuscular nature of light or electromagnetic field. The following derivation is only valid for plane waves.

It has been long known that electromagnetic energy is carried by photon, each of which is associated with a packet of energy given by $\hbar\omega$ (first discovered by Planck [24]). It is also well known that a photon has momentum given by $\hbar k$.[2] Assuming that there are photons, with density of N photons per unit volume streaming through space at the velocity of light c. Then the power

[2]It was de Broglie who posited that an electron is both a wave and a particle, the famous particle-wave duality picture, and its momentum is related to $\hbar k$ [88], where k is the wavenumber of the pertinent wave associated with the electron.

flow associated with these streaming photons is given by[3]

$$\mathbf{E} \times \mathbf{H} = \hbar\omega N c\hat{z} \tag{10.2.2}$$

Assuming that the plane wave is propagating in the z direction. Using $k = \omega/c$, we can rewrite the above more suggestively as

$$\mathbf{E} \times \mathbf{H} = \hbar k N c^2 \hat{z} \tag{10.2.3}$$

where $k = \omega/c$. Assuming that each photon has a momentum given by $\hbar k$,[4] we can relate the momentum density vector to be

$$\mathbf{G} = \hbar k N \hat{z} \tag{10.2.4}$$

which has a unit of momentum per unit volume. Then from the above, we deduce that

$$\mathbf{E} \times \mathbf{H} = \mathbf{G}c^2 = \frac{1}{\mu\varepsilon}\mathbf{G} \tag{10.2.5}$$

Or the above can be rewritten as[5]

$$\mathbf{G} = \mathbf{D} \times \mathbf{B} \tag{10.2.6}$$

where $\mathbf{D} = \varepsilon\mathbf{E}$, and $\mathbf{B} = \mu\mathbf{H}$.[6]

10.3 Complex Poynting's Theorem and Lossless Conditions

10.3.1 Complex Poynting's Theorem

It has been previously shown that the vector $\mathbf{E}(\mathbf{r}, t) \times \mathbf{H}(\mathbf{r}, t)$ has a dimension of watts/m^2 which is that of power density. Therefore, it is associated with the direction of power flow [34, 49]. As has been shown for time-harmonic field, a time average of this vector can be defined as

$$\langle \mathbf{E}(\mathbf{r}, t) \times \mathbf{H}(\mathbf{r}, t) \rangle = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{E}(\mathbf{r}, t) \times \mathbf{H}(\mathbf{r}, t)\, dt. \tag{10.3.1}$$

Given time harmonic fields $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{H}(\mathbf{r}, t)$, whose phasors are $\mathbf{E}(\mathbf{r}, \omega)$ and $\mathbf{H}(\mathbf{r}, \omega)$, respectively, we can show that for a time-harmonic field,

$$\langle \mathbf{E}(\mathbf{r}, t) \times \mathbf{H}(\mathbf{r}, t) \rangle = \frac{1}{2} \Re e\{\mathbf{E}(\mathbf{r}, \omega) \times \mathbf{H}^*(\mathbf{r}, \omega)\}. \tag{10.3.2}$$

[3] Again, we have to recalibrate our thinking. Usually, in classical thinking, we think of a moving particle having energy given by $\frac{1}{2}mv^2$, but photons are massless particles zipping around at the speed of light.

[4] Again, we have to reset our thinking as a photon is a massless particle behaving like a wave.

[5] It will be interesting to show that the units on both sides of the equation are the same. The unit of Planck constant is joule-second.

[6] The author is indebted to Wei SHA for this simple derivation.

Here, the vector $\mathbf{E}(\mathbf{r},\omega)\times\mathbf{H}^*(\mathbf{r},\omega)$, as previously discussed, is also known as the complex Poynting's vector. We define the instantaneous Poynting's vector to be

$$\mathbf{S}(\mathbf{r},t) = \mathbf{E}(\mathbf{r},t)\times\mathbf{H}(\mathbf{r},t) \tag{10.3.3}$$

and the complex Poynting's vector to be[7]

$$\underset{\sim}{\mathbf{S}}(\mathbf{r},\omega) = \mathbf{E}(\mathbf{r},\omega)\times\mathbf{H}^*(\mathbf{r},\omega) \tag{10.3.4}$$

Then for time-harmonic fields,

$$\langle\mathbf{S}(\mathbf{r},t)\rangle = \frac{1}{2}\Re e\left\{\underset{\sim}{\mathbf{S}}(\mathbf{r},\omega)\right\} \tag{10.3.5}$$

The above is often the source of confusion in the definition of Poynting's vector, because $\underset{\sim}{\mathbf{S}}$ is not the phasor representation of $\mathbf{S}$. Also, $\mathbf{S}$ is not time-harmonic and hence, does not have a phasor representation.

Since the above definition of complex Poynting's vector has the dimension of power density, we will study its conservative property. To do so, we take its divergence and use the appropriate vector identity to obtain[8]

$$\nabla\cdot(\mathbf{E}\times\mathbf{H}^*) = \mathbf{H}^*\cdot\nabla\times\mathbf{E} - \mathbf{E}\cdot\nabla\times\mathbf{H}^*. \tag{10.3.6}$$

Next, using Maxwell's equations for $\nabla\times\mathbf{E}$ and $\nabla\times\mathbf{H}^*$, namely

$$\nabla\times\mathbf{E} = -j\omega\mathbf{B} \tag{10.3.7}$$

$$\nabla\times\mathbf{H}^* = -j\omega\mathbf{D}^* + \mathbf{J}^* \tag{10.3.8}$$

and the constitutive relations for anisotropic media that

$$\mathbf{B} = \overline{\boldsymbol{\mu}}\cdot\mathbf{H}, \quad \mathbf{D}^* = \overline{\boldsymbol{\varepsilon}}^*\cdot\mathbf{E}^* \tag{10.3.9}$$

we have

$$\nabla\cdot(\mathbf{E}\times\mathbf{H}^*) = -j\omega\,\mathbf{H}^*\cdot\mathbf{B} + j\omega\,\mathbf{E}\cdot\mathbf{D}^* - \mathbf{E}\cdot\mathbf{J}^* \tag{10.3.10}$$

$$= -j\omega\,\mathbf{H}^*\cdot\overline{\boldsymbol{\mu}}\cdot\mathbf{H} + j\omega\,\mathbf{E}\cdot\overline{\boldsymbol{\varepsilon}}^*\cdot\mathbf{E}^* - \mathbf{E}\cdot\mathbf{J}^*. \tag{10.3.11}$$

The above is also known as the complex Poynting's theorem. It can also be written in an integral form using Gauss' divergence theorem, namely,

$$\int_S d\mathbf{S}\cdot(\mathbf{E}\times\mathbf{H}^*) = -j\omega\int_V dV(\mathbf{H}^*\cdot\overline{\boldsymbol{\mu}}\cdot\mathbf{H} - \mathbf{E}\cdot\overline{\boldsymbol{\varepsilon}}^*\cdot\mathbf{E}^*) - \int_V dV\mathbf{E}\cdot\mathbf{J}^*. \tag{10.3.12}$$

where S is the surface bounding the volume V.

[7] They have the same unit as the phasor of a field has the same unit as the field in the time domain.

[8] The product rule for derivative will be used, and we will drop the argument $\mathbf{r},\omega$ for the phasors in our discussion next as they will be implied.

10.3.2 Lossless Conditions

For a region V that is lossless and source-free, $\mathbf{J} = 0$. There should be no net time-averaged power-flow out of or into this region V. Therefore,

$$\Re e \int_S d\mathbf{S} \cdot (\mathbf{E} \times \mathbf{H}^*) = 0, \tag{10.3.13}$$

Because of the above, and energy conservation, the real part of the right-hand side of (10.3.11), without the $\mathbf{E} \cdot \mathbf{J}^*$ term, must also be zero. In other words, the right-hand side of (10.3.11) should be purely imaginary with no real part. Thus

$$\int_V dV(\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H} - \mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \mathbf{E}^*) \tag{10.3.14}$$

must be a purely real quantity.

Other than the possibility that the above is zero, the general requirement for (10.3.14) to be real for arbitrary $\mathbf{E}$ and $\mathbf{H}$, is that $\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}$ and $\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \mathbf{E}^*$ are real quantities. This is only possible if $\overline{\boldsymbol{\mu}}$ is hermitian.[9] Therefore, the conditions for anisotropic media to be lossless are

$$\overline{\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}^\dagger, \qquad \overline{\boldsymbol{\varepsilon}} = \overline{\boldsymbol{\varepsilon}}^\dagger, \tag{10.3.15}$$

requiring the permittivity and permeability tensors to be hermitian. If this is the case, (10.3.14) is always real for arbitrary $\mathbf{E}$ and $\mathbf{H}$, and (10.3.13) is true, implying a lossless region V.

Notice that for an isotropic medium, $\overline{\boldsymbol{\mu}} \to \mu$ and $\overline{\boldsymbol{\varepsilon}} \to \varepsilon$, this lossless conditions reduce simply to that $\Im m(\mu) = 0$ and $\Im m(\varepsilon) = 0$, or that μ and ε are pure real quantities. Looking back, many of the effective permittivities or dielectric constants that we have derived using the Drude-Lorentz-Sommerfeld model cannot be lossless when the friction term is nonzero. Looking at the formula for χ as given by (8.3.17), it cannot be real, and hence, it corresponds to a lossy medium. The friction can be due to the collision of the electrons with the lattice.

10.3.3 Anisotropic Medium Case

For a lossy medium which is conductive and anisotropic, we may define $\mathbf{J} = \overline{\boldsymbol{\sigma}} \cdot \mathbf{E}$ where $\overline{\boldsymbol{\sigma}}$ is a general conductivity tensor. In this case, equation (10.3.12), after combining the last two terms, may be written as

$$\int_S d\mathbf{S} \cdot (\mathbf{E} \times \mathbf{H}^*) = -j\omega \int_V dV \left[\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H} - \mathbf{E} \cdot \left(\overline{\boldsymbol{\varepsilon}}^* + \frac{j\overline{\boldsymbol{\sigma}}^*}{\omega}\right) \cdot \mathbf{E}^*\right] \tag{10.3.16}$$

$$= -j\omega \int dV[\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H} - \mathbf{E} \cdot \tilde{\overline{\boldsymbol{\varepsilon}}}^* \cdot \mathbf{E}^*], \tag{10.3.17}$$

[9] $\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}$ is real only if its complex conjugate, or conjugate transpose is itself. Using some details from matrix algebra that $(\overline{\mathbf{A}} \cdot \overline{\mathbf{B}} \cdot \overline{\mathbf{C}})^t = \overline{\mathbf{C}}^t \cdot \overline{\mathbf{B}}^t \cdot \overline{\mathbf{A}}^t$, implies that (in physics notation, the transpose of a vector is implied in a dot product) $(\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H})^\dagger = (\mathbf{H} \cdot \overline{\boldsymbol{\mu}}^* \cdot \mathbf{H}^*)^t = \mathbf{H}^* \cdot \overline{\boldsymbol{\mu}}^\dagger \cdot \mathbf{H} = \mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}$. The last equality in the above is possible only if $\overline{\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}^\dagger$ or that $\overline{\boldsymbol{\mu}}$ is hermitian.

where $\tilde{\tilde{\boldsymbol{\varepsilon}}} = \overline{\boldsymbol{\varepsilon}} - \frac{j\overline{\boldsymbol{\sigma}}}{\omega}$ which is the general complex permittivity tensor. In this manner, (10.3.17) has the same structure as the source-free Poynting's theorem. Notice here that the complex permittivity tensor $\tilde{\tilde{\boldsymbol{\varepsilon}}}$ is clearly non-hermitian corresponding to a lossy medium.

For a lossless medium without the source term, by taking the imaginary part of (10.3.12), we arrive at

$$\Im m \int_S d\mathbf{S} \cdot (\mathbf{E} \times \mathbf{H}^*) = -\omega \int_V dV (\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H} - \mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \mathbf{E}^*), \tag{10.3.18}$$

The left-hand side of the above is the reactive power coming out of the volume V, and hence, the right-hand side can be interpreted as reactive power as well. It is to be noted that $\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}$ and $\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \mathbf{E}^*$ are not to be interpreted as stored energy density when the medium is dispersive. The correct expressions for stored energy density in dispersive media will be derived later in the next section.

But, the quantity $\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}$ for lossless, dispersionless media is associated with the time-averaged energy density stored in the magnetic field, while the quantity $\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \mathbf{E}^*$ for lossless dispersionless media is associated with the time-averaged energy density stored in the electric field. Then, for lossless, dispersionless, source-free media, the right-hand side of the above can be interpreted as stored energy density. Therefore, the reactive power is proportional to the time rate of change of the difference of the time-averaged energy stored in the magnetic field and the electric field. For example, in a resonant cavity, these two stored energies are equal to each other, and the need for external reactive power is zero, just as the case of an LC tank circuit. In this case, energies just exchange between each other without external sources.

10.4 Energy Density in Dispersive Media[10]

This section is going to give us some new formulas previously unknown to most of us. The derivation is going to be laborious, but the punchline is that we need to update the energy storage formulas in electromagnetics. It turns out that the energy storage formula should be

$$\langle W_T \rangle = \frac{1}{4} \left[\frac{\partial \omega' \mu}{\partial \omega'} |\mathbf{H}|^2 + \frac{\partial \omega' \varepsilon}{\partial \omega'} |\mathbf{E}|^2 \right] \tag{10.4.1}$$

For a non-dispersive medium, μ and ε are independent of frequency, the above reverts back to a familiar expression,

$$\langle W_T \rangle = \frac{1}{4} \left[\mu |\mathbf{H}|^2 + \varepsilon |\mathbf{E}|^2 \right] \tag{10.4.2}$$

which is what we have derived before.

A dispersive medium alters our concept of what the formula energy density should be.[11] In order to derive the new formula, we assume that the field has complex ω in $e^{j\omega t}$, where $\omega = \omega' - j\omega''$,

[10] The derivation in this section is complex, but worth the pain, since this knowledge was not discovered until the 1960s.

[11] The derivation here is inspired by H.A. Haus, Electromagnetic Noise and Quantum Optical Measurements [89]. Generalization to anisotropic media is given by W.C. Chew, Lectures on Theory of Microwave and Optical Waveguides [90].

rather than real ω dependence. In other words, the field is not time-harmonic anymore, but quasi-time-harmonic when ω'' is very small.

We take the divergence of the complex power for fields with such a time dependence, and let $e^{j\omega t}$ be attached to the field. So $\mathbf{E}(t)$ and $\mathbf{H}(t)$ are complex field but not exactly like phasors we have studied before since they are not truly time harmonic. In other words, we let

$$\mathbf{E}(\mathbf{r},t) = \underset{\sim}{\mathbf{E}}(\mathbf{r},\omega)e^{j\omega t}, \quad \mathbf{H}(\mathbf{r},t) = \underset{\sim}{\mathbf{H}}(\mathbf{r},\omega)e^{j\omega t} \tag{10.4.3}$$

The above, just like phasors, can be made to satisfy Maxwell's equations where a time derivative becomes $j\omega$ but this time with complex ω. We can study the quantity $\mathbf{E}(\mathbf{r},t)\times\mathbf{H}^*(\mathbf{r},t)$ which still has the unit of power density. In the real ω case, their time dependence will exactly cancel each other and this quantity becomes complex power again. But now, ω is complex, and the field is quasi-time-harmonic, and their time dependences do not cancel because of the complex ω. In other words, $e^{j\omega t}e^{-j\omega^* t} = e^{2\omega'' t} \neq 1$. Therefore, as before,

$$\nabla\cdot[\mathbf{E}(t)\times\mathbf{H}^*(t)] = \mathbf{H}^*(t)\cdot\nabla\times\mathbf{E}(t) - \mathbf{E}(t)\cdot\nabla\times\mathbf{H}^*(t) \tag{10.4.4}$$

where the quantities are still time dependent, whereas in the lossless case, the right-hand side would be time independent. Maxwell's equations for this quasi-time-harmonic fields, when ω is complex, become

$$\nabla\times\mathbf{E} = -j\omega\mathbf{B} \tag{10.4.5}$$

$$\nabla\times\mathbf{H}^* = -j\omega^*\mathbf{D}^* + \mathbf{J}^* \tag{10.4.6}$$

Using the above, in (10.4.4), after using $\mathbf{B} = \mu\mathbf{H}, \quad \mathbf{D} = \varepsilon\mathbf{E}$, we arrive at

$$\nabla\cdot[\mathbf{E}(t)\times\mathbf{H}^*(t)] = -\mathbf{H}^*(t)\cdot j\omega\mu\mathbf{H}(t) + \mathbf{E}(t)\cdot j\omega^*\varepsilon^*\mathbf{E}^*(t) \tag{10.4.7}$$

where Maxwell's equations have been used to substitute for $\nabla\times\mathbf{E}(t)$ and $\nabla\times\mathbf{H}^*(t)$. The space dependence of the field is implied, and we assume a source-free medium so that $\mathbf{J} = 0$.

If $\mathbf{E}(t)\sim e^{j\omega t}$, then due to ω being complex, $\mathbf{H}^*(t)\sim e^{-j\omega^* t}$. Then the term like $\mathbf{E}(t)\times\mathbf{H}^*(t)$ is not truly time independent but becomes

$$\mathbf{E}(t)\times\mathbf{H}^*(t) \sim e^{j(\omega-\omega^*)t} = e^{2\omega'' t} \tag{10.4.8}$$

And each of the terms above will have similar time dependence. Writing (10.4.4) more explicitly, by letting $\omega = \omega' - j\omega''$, we have

$$\nabla\cdot[\mathbf{E}(t)\times\mathbf{H}^*(t)] = -j(\omega' - j\omega'')\mu(\omega)|\mathbf{H}(t)|^2 + j(\omega' + j\omega'')\varepsilon^*(\omega)|\mathbf{E}(t)|^2 \tag{10.4.9}$$

So far, everything is exact, and no approximation has been made. To simplify the complicated math here, we make some approximations!

Assuming that $\omega'' \ll \omega'$, or that the field is quasi-time-harmonic, after using Taylor series approximation, we have

$$\mu(\omega' - j\omega'') \cong \mu(\omega') - j\omega''\frac{\partial\mu(\omega')}{\partial\omega'}, \quad \varepsilon(\omega' - j\omega'') \cong \varepsilon(\omega') - j\omega''\frac{\partial\varepsilon(\omega')}{\partial\omega'} \tag{10.4.10}$$

Using (10.4.10) in (10.4.9), and collecting terms of the same order, and ignoring $(\omega'')^2$ terms, yields[12]

$$\begin{aligned}\nabla\cdot[\mathbf{E}(t)\times\mathbf{H}^*(t)] \cong & -j\omega'\mu(\omega')|\mathbf{H}(t)|^2+j\omega'\varepsilon^*(\omega')|\mathbf{E}(t)|^2\\ & -\omega''\mu(\omega')|\mathbf{H}(t)|^2-\omega'\omega''\frac{\partial\mu}{\partial\omega'}|\mathbf{H}(t)|^2\\ & -\omega''\varepsilon^*(\omega')|\mathbf{E}(t)|^2-\omega'\omega''\frac{\partial\varepsilon^*}{\partial\omega'}|\mathbf{E}(t)|^2\end{aligned}\tag{10.4.11}$$

The above can be rewritten as

$$\begin{aligned}\nabla\cdot[\mathbf{E}(t)\times\mathbf{H}^*(t)] \cong & -j\omega'\left[\mu(\omega')|\mathbf{H}(t)|^2-\varepsilon^*(\omega')|\mathbf{E}(t)|^2\right]\\ & -\omega''\left[\frac{\partial\omega'\mu(\omega')}{\partial\omega'}|\mathbf{H}(t)|^2+\frac{\partial\omega'\varepsilon^*(\omega')}{\partial\omega'}|\mathbf{E}(t)|^2\right]\end{aligned}\tag{10.4.12}$$

The above approximation is extremely good when $\omega''\ll\omega'$. For a lossless medium, $\varepsilon(\omega')$ and $\mu(\omega')$ are purely real, and the first term of the right-hand side is purely imaginary while the second term is purely real. For better physical insight, in the limit when $\omega''\to 0$, for better physical insight, we take half the imaginary part of the above equation to get

$$\nabla\cdot\frac{1}{2}\Im m\left[\mathbf{E}\times\mathbf{H}^*\right]\cong-\omega'\left[\frac{1}{2}\mu|\mathbf{H}|^2-\frac{1}{2}\varepsilon|\mathbf{E}|^2\right]\tag{10.4.13}$$

Now, the left-hand side and right-hand side of the above now can be interpreted as reactive power, something we have learnt before in complex Poynting's theorem.

When half the real part of (10.4.12) is taken, we obtain some new terms,

$$\nabla\cdot\frac{1}{2}\Re e[\mathbf{E}\times\mathbf{H}^*]=-\frac{\omega''}{2}\left[\frac{\partial\omega'\mu}{\partial\omega'}|\mathbf{H}|^2+\frac{\partial\omega'\varepsilon}{\partial\omega'}|\mathbf{E}|^2\right]\tag{10.4.14}$$

The left-hand side of the above has the physical meaning of time-average power density when $\omega''\to 0$, or in the time-harmonic limit. Since the right-hand side has time dependence of $e^{2\omega''t}$, when $\omega''\neq 0$, it can be written as

$$\nabla\cdot\frac{1}{2}\Re e[\mathbf{E}\times\mathbf{H}^*]\cong-\frac{\partial}{\partial t}\frac{1}{4}\left[\frac{\partial\omega'\mu}{\partial\omega'}|\mathbf{H}|^2+\frac{\partial\omega'\varepsilon}{\partial\omega'}|\mathbf{E}|^2\right]=-\frac{\partial}{\partial t}\langle W_T\rangle\tag{10.4.15}$$

The above is a restatement of that for a weakly time-harmonic system, the divergence of the time-average power density on the left-hand side is proportional to the time derivative of the store energy on the right-hand side.[13] This has the same physical meaning as the current continuity equation which is a statement of charge conservation. Therefore, we reproduce the formulas at the beginning of this section: the time-average stored energy density can be identified as

$$\langle W_T\rangle=\frac{1}{4}\left[\frac{\partial\omega'\mu}{\partial\omega'}|\mathbf{H}|^2+\frac{\partial\omega'\varepsilon}{\partial\omega'}|\mathbf{E}|^2\right]\tag{10.4.16}$$

[12]This is the general technique of perturbation expansion [48].

[13]Even though we say that this formula is approximately correct, to use the parlance of perturbation theory, it is correct up to leading order. Viz., it is exact when $\omega''=0$.

Again, for a non-dispersive medium, μ and ε are independent of frequency, the above reverts back to a familiar expression,

$$\langle W_T \rangle = \frac{1}{4}\left[\mu|\mathbf{H}|^2 + \varepsilon|\mathbf{E}|^2\right] \tag{10.4.17}$$

which is what we have derived before.

In the above analysis, we have used a quasi-time-harmonic signal with $\exp(j\omega t)$ dependence. In the limit when $\omega'' \to 0$, this signal reverts back to a time-harmonic signal, and to our usual interpretation of complex power. However, by assuming the frequency ω to have a very small imaginary part ω'', it forces the stored energy to grow very slightly, and hence, power has to be supplied to maintain the growth of this stored energy. By so doing, and use of energy conservation, it allows us to identify the expression for energy density for a dispersive medium. These expressions for energy density were not discovered until 1960 by Brillouin [91], as energy density times group velocity should be power flow. More discussion on this topic can be found in Jackson [49].

It is to be noted that if the same analysis is used to study the energy storage in a capacitor or an inductor, the energy storage formulas have to be altered accordingly if the capacitor or inductor is frequency dependent.

Exercises for Lecture 10

Problem 10-1:

(i) Derive (10.2.6) of the lecture notes.

(ii) Explain why the time average of the instantaneous Poynting's vector is half the real part of the complex Poynting's vector.

(iii) What is the conditions on the permittivity and permeability tensors for them to describe a lossless medium.

(iv) Derive equation (10.4.12) of the lecture notes.

Problem 10-2: Explain how the quarter-wave plate and the half-wave plate work. What kind of medium is needed to make it work? (This knowledge can be found in many textbooks or the Internet. A heuristic explanation suffices.)

Problem 10-3: Show that the units on the left-hand side and the right-hand side of (10.2.6) are equal to each other.

Chapter 11

Uniqueness Theorem

The uniqueness of a solution to a linear system of equations is an important concept in mathematics. Likewise, linear systems described by ordinary differential equation, partial differential equation, and matrix equations will have unique solutions under the prescribed boundary conditions and the driving source terms. This is the way we solve a boundary value problem. But uniqueness of a boundary value problem is not always guaranteed as we shall see unless additional conditions are stipulated. This issue is discussed in many math books and linear algebra books [92, 83]. The proof of uniqueness for Laplace and Poisson equations are given in [57, 33] which is slightly different from the electrodynamic case.

To study uniqueness with Maxwell's equations is rather involved. But solving Maxwell's equations is analogous (or "homomorphic") to solving a system of linear algebra equations. Therefore, the uniqueness of Maxwell's equations is related to the uniqueness of linear algebra equations. You may not see it now, but eventually, you will get the epiphany that they are the same!

Just imagine how bizzare it would be if there are more than one possible solutions. One has to determine which is the real solution. To quote Star Trek, we need to know who the real McCoy is![1]

11.1 The Difference Solutions to Source-Free Maxwell's Equations

In this section, we will prove uniqueness theorem for electrodynamic problems under the prescribed boundary condition with unique sources in the system [53, 37, 68, 34, 90]. This is important, as when we solve Maxwell's equations, we are solving a set of partial differential equations as a boundary value problem with prescribed boundary conditions. We like to know the conditions under which such a problem has a unique solution.

First, let us assume that uniqueness is not guaranteed; there exist two solutions Maxwell's equations in the presence of one set of common impressed sources $\mathbf{J}_i$ and $\mathbf{M}_i$.[2] Namely, these two

[1] This phrase was made popular to the baby-boom generation, or the Trekkies by Star Trek. It actually refers to an African American inventor.

[2] It is not clear when the useful concept of impressed sources were first used in electromagnetics even though

solutions are $\mathbf{E}^a$, $\mathbf{H}^a$, $\mathbf{E}^b$, $\mathbf{H}^b$. Both of them satisfy Maxwell's equations and the same boundary conditions, and also with a set of common impressed sources. What additional conditions do we need to impose so that $\mathbf{E}^a = \mathbf{E}^b$ and $\mathbf{H}^a = \mathbf{H}^b$?

Uniqueness for Maxwell's equations is only easily proved for linear Maxwell's equations, which is "homomorphic" to a system of linear algebraic equations. Therefore, to study the uniqueness theorem, we consider general linear anisotropic inhomogeneous media, where the tensors $\overline{\boldsymbol{\mu}}$ and $\overline{\boldsymbol{\varepsilon}}$ can be complex so that lossy media can be included. In the frequency domain, let us assume two possible solutions with only one given set of sources $\mathbf{J}_i$ and $\mathbf{M}_i$. It follows that

$$\nabla \times \mathbf{E}^a = -j\omega\overline{\boldsymbol{\mu}} \cdot \mathbf{H}^a - \mathbf{M}_i \tag{11.1.1}$$

$$\nabla \times \mathbf{E}^b = -j\omega\overline{\boldsymbol{\mu}} \cdot \mathbf{H}^b - \mathbf{M}_i \tag{11.1.2}$$

$$\nabla \times \mathbf{H}^a = j\omega\overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}^a + \mathbf{J}_i \tag{11.1.3}$$

$$\nabla \times \mathbf{H}^b = j\omega\overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}^b + \mathbf{J}_i \tag{11.1.4}$$

By taking the difference of these two solutions, we have

$$\nabla \times (\mathbf{E}^a - \mathbf{E}^b) = -j\omega\overline{\boldsymbol{\mu}} \cdot (\mathbf{H}^a - \mathbf{H}^b) \tag{11.1.5}$$

$$\nabla \times (\mathbf{H}^a - \mathbf{H}^b) = j\omega\overline{\boldsymbol{\varepsilon}} \cdot (\mathbf{E}^a - \mathbf{E}^b) \tag{11.1.6}$$

Or alternatively, defining $\delta\mathbf{E} = \mathbf{E}^a - \mathbf{E}^b$ and $\delta\mathbf{H} = \mathbf{H}^a - \mathbf{H}^b$, we have

$$\nabla \times \delta\mathbf{E} = -j\omega\overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} \tag{11.1.7}$$

$$\nabla \times \delta\mathbf{H} = j\omega\overline{\boldsymbol{\varepsilon}} \cdot \delta\mathbf{E} \tag{11.1.8}$$

The difference solutions, $\delta\mathbf{E}$ and $\delta\mathbf{H}$, satisfy the original source-free Maxwell's equations. Source-free here implies that we are looking at the homogeneous solutions of the pertinent partial differential equations constituted by (11.1.7) and (11.1.8). They are also analogous to finding the null-space solution of a matrix equation in linear algebra.

To prove uniqueness, we would like to find a simplifying expression for $\nabla \cdot (\delta\mathbf{E} \times \delta\mathbf{H}^*)$. By using the product rule for divergence operator, and the scalar-triple product identity, it can be shown that

$$\nabla \cdot (\delta\mathbf{E} \times \delta\mathbf{H}^*) = \delta\mathbf{H}^* \cdot \nabla \times \delta\mathbf{E} - \delta\mathbf{E} \cdot \nabla \times \delta\mathbf{H}^* \tag{11.1.9}$$

We need to simplify the right-hand side of the above with the goal of proving the uniqueness theorem. Then by taking the left dot product of $\delta\mathbf{H}^*$ with (11.1.7), and then the left dot product of $\delta\mathbf{E}$ with the complex conjugation of (11.1.8), we obtain

$$\begin{aligned} \delta\mathbf{H}^* \cdot \nabla \times \delta\mathbf{E} &= -j\omega\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} \\ \delta\mathbf{E} \cdot \nabla \times \delta\mathbf{H}^* &= -j\omega\delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^* \end{aligned} \tag{11.1.10}$$

Now, taking the difference of the above, we get

$$\begin{aligned} \delta\mathbf{H}^* \cdot \nabla \times \delta\mathbf{E} - \delta\mathbf{E} \cdot \nabla \times \delta\mathbf{H}^* &= \nabla \cdot (\delta\mathbf{E} \times \delta\mathbf{H}^*) \\ &= -j\omega\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} + j\omega\delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^* \end{aligned} \tag{11.1.11}$$

it was used in [93] in 1936. These are immutable sources that cannot be changed by the environment in which they are immersed. They are like current and voltage sources in circuit theory that are immutable irrespective of environmental loading, e.g., by changing the circuit elements in the circuit.

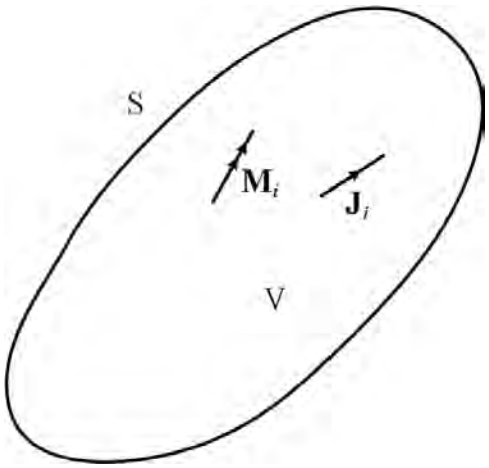


Figure 11.1: Geometry for proving the uniqueness theorem. We like to know the requisite boundary conditions on S plus the type of media inside V in order to guarantee the uniqueness of the solution in V.

Our goal is to find the conditions under which $\delta\mathbf{H}$ and $\delta\mathbf{E}$ are both zero, which will guarantee uniqueness of the solution. Next, we integrate the above equation (11.1.11) over a volume V bounded by a surface S as shown in Figure 11.1. After making use of Gauss' divergence theorem, we arrive at

$$\iiint_V \nabla \cdot (\delta\mathbf{E} \times \delta\mathbf{H}^*) dV = \oiint_S (\delta\mathbf{E} \times \delta\mathbf{H}^*) \cdot d\mathbf{S} \tag{11.1.12}$$

or after equating with the right-hand side of (11.1.11), we have that

$$\oiint_S (\delta\mathbf{E} \times \delta\mathbf{H}^*) \cdot d\mathbf{S} = \iiint_V [-j\omega\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} + j\omega\delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^*] dV \tag{11.1.13}$$

And next, we would like to know the kind of boundary conditions that would make the left-hand side equal to zero such that we can make a statement on $\delta\mathbf{H}$ and $\delta\mathbf{E}$ on the right-hand side. It is seen that the surface integral on the left-hand side will be zero if:[3]

1. If $\hat{n} \times \mathbf{E}$ is specified over S for the two possible solutions, so that $\hat{n} \times \mathbf{E}_a = \hat{n} \times \mathbf{E}_b$ on S. Then $\hat{n} \times \delta\mathbf{E} = 0$, which is the PEC boundary condition for $\delta\mathbf{E}$, and then[4]

$\oiint_S (\delta\mathbf{E} \times \delta\mathbf{H}^*) \cdot \hat{n} dS = \oiint_S (\hat{n} \times \delta\mathbf{E}) \cdot \delta\mathbf{H}^* dS = 0.$

[3]In the following, please be reminded that PEC stands for "perfect electric conductor", while PMC stands for "perfect magnetic conductor". PMC is the dual of PEC. Also, a fourth case of impedance boundary condition is possible, which is beyond the scope of this course. Interested readers may consult Chew, Theory of Microwave and Optical Waveguides [90].

[4]We use the vector identity (also called the scalar-triple product) that $\mathbf{a} \cdot (\mathbf{b} \times c) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a})$. Also, from Section 1.3.3, $d\mathbf{S} = \hat{n} dS$.

2. If $\hat{n} \times \mathbf{H}$ is specified over S for the two possible solutions, so that $\hat{n} \times \mathbf{H}_a = \hat{n} \times \mathbf{H}_b$ on S. Then $\hat{n} \times \delta\mathbf{H} = 0$, which is the PMC boundary condition for $\delta\mathbf{H}$, and then
$\oiint_S (\delta\mathbf{E} \times \delta\mathbf{H}^*) \cdot \hat{n} dS = -\oiint_S (\hat{n} \times \delta\mathbf{H}^*) \cdot \delta\mathbf{E} dS = 0.$

3. Let the surface S be divided into two mutually exclusive surfaces S_1 and S_2.[5] If $\hat{n} \times \mathbf{E}$ is specified over S_1, and $\hat{n} \times \mathbf{H}$ is specified over S_2. Then $\hat{n} \times \delta\mathbf{E} = 0$ (PEC boundary condition) on S_1, and $\hat{n} \times \delta\mathbf{H} = 0$ (PMC boundary condition) on S_2. Therefore, the left-hand side becomes
$\oiint_S (\delta\mathbf{E} \times \delta\mathbf{H}^*) \cdot \hat{n} dS = \iint_{S_1} + \iint_{S_2} = \iint_{S_1} (\hat{n} \times \delta\mathbf{E}) \cdot \delta\mathbf{H}^* dS$
$- \iint_{S_2} (\hat{n} \times \delta\mathbf{H}^*) \cdot \delta\mathbf{E} dS = 0.$

Thus, under the above three scenarios, the left-hand side of (11.1.13) is zero, and then the right-hand side of (11.1.13) becomes

$$\iiint_V [-j\omega\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} + j\omega\delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^*] dV = 0 \tag{11.1.14}$$

For lossless media, $\overline{\boldsymbol{\mu}}$ and $\overline{\boldsymbol{\varepsilon}}$ are hermitian tensors (or matrices[6]), then it can be seen, using the properties of hermitian matrices or tensors, that $\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H}$ and $\delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^*$ are purely real. Taking the imaginary part of the above equation yields

$$\iiint_V [-\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} + \delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^*] dV = 0 \tag{11.1.15}$$

The above two terms correspond to stored magnetic field energy and stored electric field energy in the difference solutions $\delta\mathbf{H}$ and $\delta\mathbf{E}$, respectively. The above being zero does not imply that $\delta\mathbf{H}$ and $\delta\mathbf{E}$ are zero since they can be negative of each other.

For resonant solutions, the stored electric energy and stored magnetic energy can balance (or cancel) each other implying that $\delta\mathbf{H}$ and $\delta\mathbf{E}$ need not be zero when the above is zero. The above resonant solutions are those of the difference solutions satisfying PEC or PMC boundary condition or mixture thereof. Also, they are the solutions of the source-free Maxwell's equations (11.1.7).[7] Clearly, $\delta\mathbf{H}$ and $\delta\mathbf{E}$ need not be zero, even though (11.1.15) is zero. This happens when we encounter solutions that are the resonant modes of the volume V bounded by the surface S.

11.2 Conditions for Uniqueness

Uniqueness can only be guaranteed if the medium is lossy as shall be shown later. It is also guaranteed if lossy impedance boundary conditions are imposed.[8] First we begin with the isotropic case which is simple.

[5] In math parlance, $S_1 \cup S_2 = S$.
[6] Tensors are a special kind of matrices that relate two physical vectors.
[7] They are also the homogeneous solution of a partial differential equation.
[8] See Chew, Theory of Microwave and Optical Waveguides.

11.2.1 Isotropic Case

It is easier to see this for lossy isotropic media. Then (11.1.14) simplifies to

$$\iiint_V [-j\omega\mu|\delta\mathbf{H}|^2 + j\omega\varepsilon^*|\delta\mathbf{E}|^2]dV = 0 \tag{11.2.1}$$

For isotropic lossy media, $\mu = \mu' - j\mu''$ and $\varepsilon = \varepsilon' - j\varepsilon''$. Taking the real part of the above, we have from (11.2.1) that, "miraculously",

$$\iiint_V [-\omega\mu''|\delta\mathbf{H}|^2 - \omega\varepsilon''|\delta\mathbf{E}|^2]dV = 0 \tag{11.2.2}$$

Now the two terms in the integrand are of the same sign. Moreover, $|\delta\mathbf{H}|^2$ and $|\delta\mathbf{E}|^2$ are always positive definite. Thus the integrand in the above is always negative definite, the integral can be zero only if

$$\delta\mathbf{E} = 0, \quad \delta\mathbf{H} = 0 \tag{11.2.3}$$

everywhere in V, implying that $\mathbf{E}_a = \mathbf{E}_b$, and that $\mathbf{H}_a = \mathbf{H}_b$ everywhere in V. Consequently, it is seen that uniqueness is guaranteed only if the medium is lossy.

The physical reason is that when the medium is lossy, a homogeneous solution (also called a natural solution, a resonant solution, or a modal solution) which is pure time-harmonic solution cannot exist due to loss. The modes or the natural solutions, which are the source-free solutions of Maxwell's equations, are decaying sinusoids. But when we express equations (11.1.1) to (11.1.4) in the frequency domain, we are seeking time-harmonic solutions for which ω is real. Hence, (11.2.3) is true in order for (11.2.2) to be true.

Notice that the same conclusion can be drawn if we make μ'' and ε'' negative. This corresponds to active media, and uniqueness can be guaranteed for a time-harmonic solution if μ'' and ε'' are of the same sign. Again, here no natural time-harmonic solution can exist, since the resonant solution or the homogeneous solution is a growing sinusoid.

Therefore, uniqueness is guaranteed for active or passive media. However, if the medium is a mixed of active and passive media, uniqueness is not guaranteed again as the equation (11.2.2) can be satisfied with $\delta\mathbf{H}$ and $\delta\mathbf{E}$ being not zero if μ'' and ε'' are of the opposite signs.

11.2.2 General Anisotropic Case

The proof for general anisotropic media is more involved. For the lossless anisotropic media, we see that (11.1.14) is purely imaginary. However, when the medium is lossy, this same equation is not purely imaginary, and will have a real part. Hence, we need to find the real part of (11.1.14) for the general lossy case.

About taking the Real and Imaginary Parts of a Complicated Expression

To this end, we digress on taking the real and imaginary parts of a complicated expression. Here, we need to find the complex conjugate[9] of the integrand of (11.1.14), which is a scalar, and add it

[9] Also called hermitian conjugate.

to itself to get its real part. To this end, we will find the conjugate of its integrand which is also a scalar number.

First, the complex conjugate of the first scalar term in the integrand of (11.1.14) is[10]

$$(-j\omega\delta\mathbf{H}^*\cdot\overline{\boldsymbol{\mu}}\cdot\delta\mathbf{H})^* = j\omega\delta\mathbf{H}\cdot\overline{\boldsymbol{\mu}}^*\cdot\delta\mathbf{H}^* = j\omega\delta\mathbf{H}^*\cdot\overline{\boldsymbol{\mu}}^\dagger\cdot\delta\mathbf{H} \tag{11.2.4}$$

The last equality follows because we are just taking the transpose of a scalar number. Similarly, the complex conjugate of the second scalar term in the same integrand is

$$(j\omega\delta\mathbf{E}\cdot\overline{\boldsymbol{\varepsilon}}^*\cdot\delta\mathbf{E}^*)^* = -j\omega\delta\mathbf{E}^*\cdot\overline{\boldsymbol{\varepsilon}}^\dagger\cdot\delta\mathbf{E} \tag{11.2.5}$$

But

$$j\omega\delta\mathbf{E}\cdot\overline{\boldsymbol{\varepsilon}}^*\cdot\delta\mathbf{E}^* = j\omega\delta\mathbf{E}^*\cdot\overline{\boldsymbol{\varepsilon}}^\dagger\cdot\delta\mathbf{E} \tag{11.2.6}$$

The above gives us the complex conjugate of the scalar quantity (11.1.14) and adding it to itself, we have

$$\iiint_V [-j\omega\delta\mathbf{H}^*\cdot(\overline{\boldsymbol{\mu}}-\overline{\boldsymbol{\mu}}^\dagger)\cdot\delta\mathbf{H} - j\omega\delta\mathbf{E}^*\cdot(\overline{\boldsymbol{\varepsilon}}-\overline{\boldsymbol{\varepsilon}}^\dagger)\cdot\delta\mathbf{E}]dV = 0 \tag{11.2.7}$$

For lossy media, $-j(\overline{\boldsymbol{\mu}}-\overline{\boldsymbol{\mu}}^\dagger)$ and $-j(\overline{\boldsymbol{\varepsilon}}-\overline{\boldsymbol{\varepsilon}}^\dagger)$ are hermitian positive matrices. Hence the integrand is always positive definite, and the above equation cannot be satisfied unless $\delta\mathbf{H} = \delta\mathbf{E} = 0$ everywhere in V. Uniqueness is thus guaranteed in a lossy anisotropic medium.

Similar statement can be made for the anisotropic case if the medium is active. Then the integrand is positive definite, and the above equation cannot be satisfied unless $\delta\mathbf{H} = \delta\mathbf{E} = 0$ everywhere in V, thereby proving that uniqueness is satisfied.

11.3 Hind Sight Using Linear Algebra

The proof of uniqueness for Maxwell's equations is very similar to the proof of uniqueness for a matrix equation [83]. The two problems are "homomorphic" to each other. As you will see, the proof using linear algebra is a lot simpler due to the simplicity of notations. To see this, consider a linear algebraic equation

$$\overline{\mathbf{A}}\cdot\mathbf{x} = \mathbf{b} \tag{11.3.1}$$

If a solution to a matrix equation exists without excitation, namely, when $\mathbf{b} = 0$, then the solution to (11.3.1) is the null space solution [83], namely, $\mathbf{x} = \mathbf{x}_N$, or

$$\overline{\mathbf{A}}\cdot\mathbf{x}_N = 0 \tag{11.3.2}$$

For Maxwell's equations, $\mathbf{b}$ corresponds to the source terms. In a word, the solution in (11.3.2) is like the homogeneous solution of an ordinary differential equation or a partial differential equation

[10] To arrive at these expressions, one makes use of the matrix algebra rule that if $\overline{\mathbf{D}} = \overline{\mathbf{A}}\cdot\overline{\mathbf{B}}\cdot\overline{\mathbf{C}}$, then $\overline{\mathbf{D}}^t = \overline{\mathbf{C}}^t\cdot\overline{\mathbf{B}}^t\cdot\overline{\mathbf{A}}^t$. This is true even for non-square matrices. But for our case here, $\overline{\mathbf{A}}$ is a 1×3 row vector, and $\overline{\mathbf{C}}$ is a 3×1 column vector, and $\overline{\mathbf{B}}$ is a 3×3 matrix. In vector algebra, the transpose of a vector is implied. Also, in our case here, $\overline{\mathbf{D}}$ is a scalar, and hence, its transpose is itself.

[92]. They are also analogous to the null-space solution of a matrix equation. In an enclosed region of volume V bounded by a surface S, homogeneous solutions are the resonant solutions (or the natural solutions) of this Maxwellian system. When these solutions exist, they give rise to non-uniqueness. (Note that these resonant solutions in the time domain exist for all time if the cavity is lossless.)

Also, notice that (11.1.7) and (11.1.8) for the difference solutions are Maxwell's equations without the source terms. When there are no source terms in a closed, lossless region V bounded by a surface S, only resonant solutions can exist for $\delta\mathbf{E} \neq 0$ and $\delta\mathbf{H} \neq 0$ with the relevant boundary conditions.

As previously mentioned, one way to ensure that these resonant solutions (or homogeneous solutions) are eliminated is to put in loss or gain. When loss or gain is present, then the resonant solutions are decaying sinusoids or growing sinusoids. Since we are looking for solutions in the frequency domain, or time harmonic solutions, the solutions we are seeking are on the real ω axis on the complex ω plane. Thus the non-sinusoidal (non-time-harmonic) solutions are outside the solution space: They are not part of the time-harmonic solutions (which are on the real axis) that we are looking for.

We see that the source of non-uniqueness is the homogeneous solutions or the resonant solutions of the system that persist for all time. These solutions are non-causal, and they are there in the system since the beginning of time and forever thereafter. One way to remove these resonant solutions is to set them to zero at the beginning by solving an initial value problem (IVP). However, this has to be done in the time domain. Therefore, one reason for non-uniqueness is because we are seeking the solutions in the frequency domain.

11.4 Connection to Poles of a Linear System

The output is the response to the input of a linear system. It can be represented by a transfer function $H(\omega)$ [55, 94]. If $H(\omega)$ has poles, and if the system is lossless, the poles are on the real axis. Therefore, when $\omega = \omega_{\text{pole}}$, the function $H(\omega)$ becomes undefined. In other words, one can add a constant term to the output, and the ratio between output to input is still infinity. This also is the reason for non-uniqueness of the output with respect to the input. Poles usually correspond to resonant solutions, and hence, the non-uniqueness of the solution is intimately related to the non-uniqueness of Maxwell's equations at the resonant frequencies of a structure. This is illustrated in the upper part of Figure 11.2.

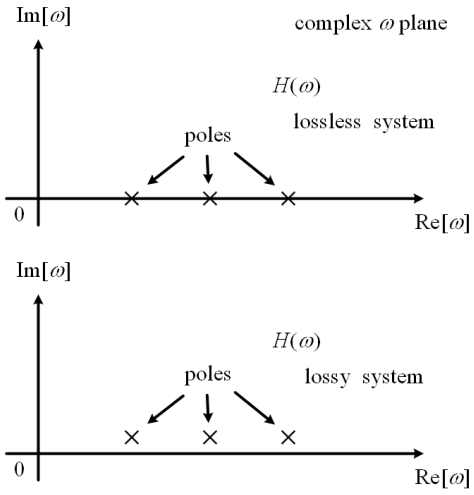


Figure 11.2: The non-uniqueness problem is intimately related to the locations of the poles of a transfer function being on the real axis, when one solves a linear system using Fourier transform technique. For a lossless system, the poles are located on the real axis, then the Fourier inverse transform is not defined since Fourier inversion contour is on the real axis. But when performing a Fourier inverse transform to obtain the solution in the time domain, then the Fourier inversion contour is undefined, and the solution cannot be uniquely determined.

If the input function is $f(t)$, with Fourier transform $F(\omega)$, then the output $y(t)$ is given by the following Fourier integral, viz.,

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega e^{j\omega t} H(\omega) F(\omega) \tag{11.4.1}$$

where the Fourier inversion integral path is on the real axis on the complex ω plane. The Fourier inversion integral above is undefined or non-unique if poles exist on the real ω axis.

However, if loss is introduced, these poles will move away from the real axis as shown in the lower part of Figure 11.2. Then the transfer function is uniquely determined for all frequencies on the real axis. In this way, the Fourier inversion integral in (11.4.1) is well defined, and uniqueness of the solution is guaranteed.

When the poles are located on the real axis yielding possibly non-unique solutions, a remedy to this problem is to use Laplace transform technique [55] when solving such problems. The Laplace transform technique allows the specification of initial values, which is similar to solving the problem as an initial value problem (IVP). As mentioned before, solving these problems as an IVP will remove non-uniqueness of the solution.

If you have problem wrapping your head around this concept, it is good to connect back to the LC tank circuit example. The transfer function $H(\omega)$ is similar to the $Y(\omega) = j\omega C + \frac{1}{j\omega L}$. The transfer function has two poles at $\omega = \pm\sqrt{LC}$. If there is no loss, then the poles are located on

the real ω axis, rendering the Fourier inversion contour undefined in (11.4.1). Hence, the solution is non-unique. However, if infinitesimal loss is introduced by adding a resistor, then the poles will migrate off the real axis making the Fourier integral in (11.4.1) well defined!

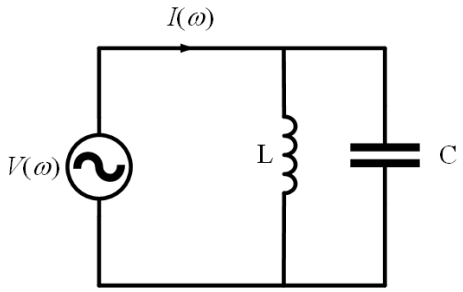


Figure 11.3: The transfer function of the LC tank circuit is $H(\omega) = Z(\omega)$ where the input is $I(\omega)$ and the output is $V(\omega) = H(\omega)I(\omega)$. One can show that the transfer function $H(\omega)$ has poles on the real ω axis, implying that the Fourier inverse transform in (11.4.1) where the integration is on the real ω axis does not exist. But when a small resistor R is added in series with the inductor however, the poles are shifted off from the real axis making the Fourier transform in (11.4.1) well defined.

11.5 Radiation from Antenna Sources and Radiation Condition[11]

The above uniqueness theorem guarantees that if we have some antennas with prescribed current sources on them, the radiated field from these antennas are unique under certain conditions. To see how this can come about, we first study the radiation of sources into a region V bounded by a large surface $S_{\text{large}} \to S_{\text{inf}}$ as shown in Figure 11.4 [37].

When we solve a boundary value problem, $\hat{n} \times \mathbf{E}$ or $\hat{n} \times \mathbf{H}$ are specified on the surface at S_{large}. When the volume V bounded by S_{large} is large, the problem can have many resonant solutions. In fact, the region will be replete with resonant solutions as one makes S_{large} become very large. When we pick an operating frequency ω, the likelihood of ω coinciding with a resonant solution is high.

[11] May be skipped on first reading.

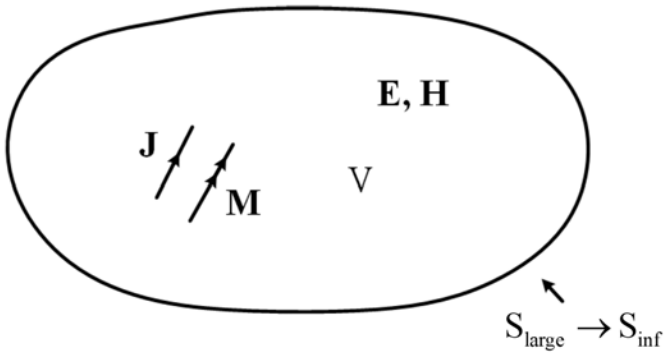


Figure 11.4: When $S_{\text{large}} \to S_{\text{inf}}$, the number of the resonant solution within the surface S becomes infinitely large. The chance of the operating frequency ω coinciding with the resonant frequency is 1. However, the solution can be made unique by assuming an infinitesimal loss. Therefore, the solution for antenna radiation in an infinite space can be made unique by imposing the Sommerfeld radiation condition. That is we assume that the radiation wave travels to infinity but never to return. This is equivalent to assuming an infinitesimal loss when seeking the solution in V and later let $V \to \infty$.

To gain more insight, we look at the resonant condition of a large rectangular cavity reproduced here as[12]

$$\beta^2 = \frac{\omega^2}{c^2} = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{p\pi}{d}\right)^2 \tag{11.5.1}$$

The above is an equation of an Ewald sphere in a 3D mode space which is described by discrete points, or that the values of $\beta_x = \frac{m\pi}{a}$, $\beta_y = \frac{n\pi}{b}$, and $\beta_z = \frac{p\pi}{d}$ are discrete. Here, β_x, β_y, and β_z can be thought of the Fourier transform variables of x, y, and z. These points are known as reciprocal lattice points in solid state physics [95, 96].

We can continuously change the operating frequency ω above until the above equation is satisfied. When this happens, we encounter a resonant frequency of the cavity. At this operating frequency, the solution to Maxwell's equations inside the cavity is non-unique. As the dimensions of the cavity become large (or a, b, and d are large), then the number of ω's or resonant frequencies at which the above equation can be satisfied becomes very large. This is illustrated Figure 11.5 in 2D. Hence, the chance of the operating frequency ω coinciding with a resonant mode of the cavity is very high (with probability of almost 1) giving rise to non-uniqueness. This above argument applies to cavities of other shapes as well.

[12]This formula is usually covered in an undergraduate course in electromagnetics, and will be covered later in this course in Section 25.2.1.

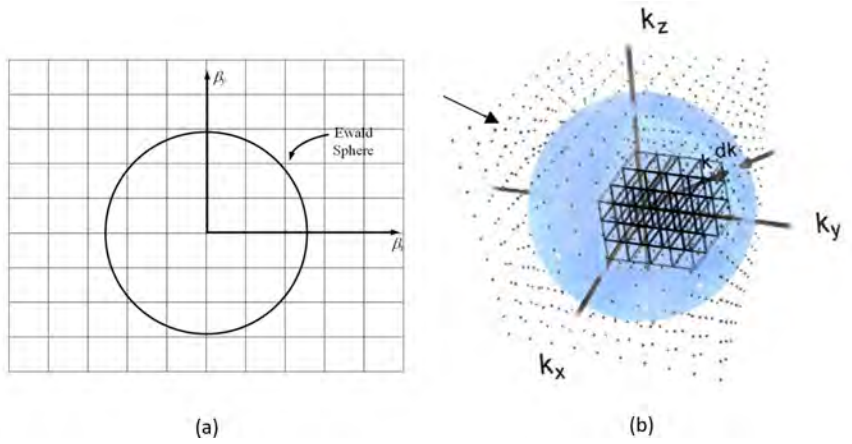


Figure 11.5: For a very large cavity, the grid spacing in the mode space (or Fourier space, also called the reciprocal lattice space [95, 96]) becomes very small. The radius of the sphere is given by β. Then the chance that the sphere surface encounters a resonant mode is very high. When this happens, the solution to the cavity problem is non-unique. (a) shows the 2D case while (b) shows the 3D case (courtesy of W.A. Doolittle, Georgia Tech). The sphere defined by the equation $\beta^2 = \beta_x^2 + \beta_y^2 + \beta_z^2$ is known as the Ewald sphere. Here, k is equivalent to β.

The way to remove these resonant solutions is to introduce an infinitesimal amount of loss in region V. Then these resonant solutions will disappear from the real ω axis, where we seek a time-harmonic solution. Now we can take S_{large} to S_{inf} which is infinitely large, and the solution will always remains unique even if the loss is infinitesimally small.

Notice that if $S_{\text{inf}} \to \infty$, the waves that leave the sources will never be reflected back because of the small amount of loss. The radiated field will just disappear into infinity. This is just what radiation loss is: power that propagates to infinity, but never to return. In fact, one way of guaranteeing the uniqueness of the solution in region V when S_{inf} is infinitely large, or that V is infinitely large is to impose the radiation condition: the waves that radiate to infinity are outgoing waves only, and never do they return. This is also called the *Sommerfeld radiation condition* [97]. Uniqueness of the field outside the sources is always guaranteed if we assume that the field radiates to infinity and never to return. This is equivalent to solving the cavity solutions with an infinitesimal loss, and then letting the size of the cavity become infinitely large.

Exercises for Lecture 11

Problem 11-1:

(i) Derive Equation (11.1.13) of the lecture notes.

(ii) Explain the conditions for uniqueness of the solutions to Maxwell's equations.

(iii) Explain why the non-uniqueness of Maxwell's equations is similar to the non-uniqueness of solving a matrix equation with a null space.

(iv) Explain why when one solves for the solutions of Maxwell's equations in a very large lossless cavity, the probability of encountering non-uniqueness is very high.

Chapter 12

Reciprocity Theorem

Reciprocity theorem is one of the most important theorems in electromagnetics. With it we can develop physical intuition and sanity check to ascertain the correctness of a certain design or experiment. Often time, a lossless reciprocal system is also time-reversible.

Reciprocity theorem is like "tit-for-tat" relationship in humans: Good-will is reciprocated with good will while ill-will is countered with ill-will. Both Confucius (551 BC–479 BC) and Jesus Christ (4 BC–AD 30) epoused the concept that, "Don't do unto others that you don't like others to do unto you." But in electromagnetics, this beautiful relationship can be expressed precisely and succinctly using mathematics. We shall see how this is done.

Figure 12.1: (Left) A depiction of Confucius from a stone fresco from the Western Han dynasty (202 BC–9 AD). The emphasis of the importance of "reciprocity" by Confucius Analects as translated by D. Hinton [98]. (Right) A portrait of Jesus that is truer to its form. Jesus teaching from the New Testament says, "Do unto others as you would have them do unto you." Luke 6:31 and Matthew 7:12 [99]. You may recall what the Old Testament says.

12.1 Mathematical Derivation

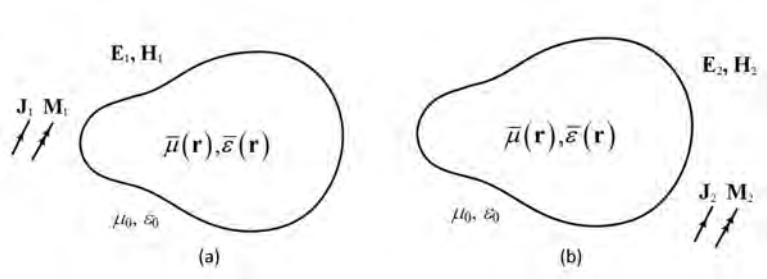


Figure 12.2: The geometry for proving reciprocity theorem. We perform two experiments on the same object or medium: (a) With sources $\mathbf{J}_1$ and $\mathbf{M}_1$ turned on, generating fields $\mathbf{E}_1$ and $\mathbf{H}_1$, and (b) With sources $\mathbf{J}_2$ and $\mathbf{M}_2$ turned on, generating fields $\mathbf{E}_2$ and $\mathbf{H}_2$. Magnetic currents, by convention, are denoted by double arrows.

Consider a general linear anisotropic inhomogeneous medium in the frequency domain where both $\overline{\boldsymbol{\mu}}(\mathbf{r})$ and $\overline{\boldsymbol{\varepsilon}}(\mathbf{r})$ are described by permeability tensor and permittivity tensor over a finite part of space as shown in Figure 12.2. In the phasor world, this representation of the medium is quite general, as it can include dispersive and conductive media. It can represent complex terrain, or complicated electronic circuit structures in circuit boards or microchips, as well as complicated antenna structures (see Figure 12.5).

We will do a Gedanken experiment[1] where a scatterer or an object is illuminated by fields from two different sets of sources which are turned on and off consecutively. This is illustrated in Figure 12.2: The geometry is not changed, but when only $\mathbf{J}_1$ and $\mathbf{M}_1$, as impressed sources, are turned on (case (a)), they generate fields $\mathbf{E}_1$ and $\mathbf{H}_1$ in this medium. On the other hand, when only $\mathbf{J}_2$ and $\mathbf{M}_2$ as impressed sources are turned on (case (b)), they generate $\mathbf{E}_2$ and $\mathbf{H}_2$ in this medium. Therefore, the pertinent equations in the frequency domain, for linear time-invariant systems, for these two cases are[2]

$$\nabla \times \mathbf{E}_1 = -j\omega\overline{\boldsymbol{\mu}} \cdot \mathbf{H}_1 - \mathbf{M}_1 \tag{12.1.1}$$

$$\nabla \times \mathbf{H}_1 = j\omega\overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_1 + \mathbf{J}_1 \tag{12.1.2}$$

$$\nabla \times \mathbf{E}_2 = -j\omega\overline{\boldsymbol{\mu}} \cdot \mathbf{H}_2 - \mathbf{M}_2 \tag{12.1.3}$$

$$\nabla \times \mathbf{H}_2 = j\omega\overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_2 + \mathbf{J}_2 \tag{12.1.4}$$

To prove reciprocity, we would like to find a simplifying expression for the divergence of the following quantity,

$$\nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2) = \mathbf{H}_2 \cdot \nabla \times \mathbf{E}_1 - \mathbf{E}_1 \cdot \nabla \cdot \mathbf{H}_2 \tag{12.1.5}$$

[1] Thought experiment in German.

[2] The current sources are impressed currents so that they are immutable, and not changed by the environment in which they are immersed [93, 53].

so that the divergence theorem can be invoked. We need to expand the right-hand side further so that reciprocity relationships can be derived. To this end, and from the above, we can show that (after left dot-multiply (12.1.1) with $\mathbf{H}_2$, and (12.1.4) with $\mathbf{E}_1$),

$$\mathbf{H}_2 \cdot \nabla \times \mathbf{E}_1 = -j\omega \mathbf{H}_2 \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}_1 - \mathbf{H}_2 \cdot \mathbf{M}_1 \tag{12.1.6}$$

$$\mathbf{E}_1 \cdot \nabla \times \mathbf{H}_2 = j\omega \mathbf{E}_1 \cdot \overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_2 + \mathbf{E}_1 \cdot \mathbf{J}_2 \tag{12.1.7}$$

Then, using the above and subtracting them, following identity in (12.1.5) and the above, we get the second equality in the following expression:

$$\begin{aligned} \nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2) &= \mathbf{H}_2 \cdot \nabla \times \mathbf{E}_1 - \mathbf{E}_1 \cdot \nabla \times \mathbf{H}_2 \\ &= -j\omega \mathbf{H}_2 \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}_1 - j\omega \mathbf{E}_1 \cdot \overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_2 - \mathbf{H}_2 \cdot \mathbf{M}_1 - \mathbf{E}_1 \cdot \mathbf{J}_2 \end{aligned} \tag{12.1.8}$$

By the same token,

$$\nabla \cdot (\mathbf{E}_2 \times \mathbf{H}_1) = -j\omega \mathbf{H}_1 \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}_2 - j\omega \mathbf{E}_2 \cdot \overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_1 - \mathbf{H}_1 \cdot \mathbf{M}_2 - \mathbf{E}_2 \cdot \mathbf{J}_1 \tag{12.1.9}$$

If one assumes that

$$\overline{\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}^t, \qquad \overline{\boldsymbol{\varepsilon}} = \overline{\boldsymbol{\varepsilon}}^t \tag{12.1.10}$$

or when the tensors are symmetric, then it follows that $\mathbf{H}_1 \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}_2 = \mathbf{H}_2 \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}_1$ and $\mathbf{E}_1 \cdot \overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_2 = \mathbf{E}_2 \cdot \overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_1$.[3]

Upon subtracting (12.1.8) and (12.1.9), many terms not involving the currents cancel each other, and one gets a simplified equation

$$\nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) = -\mathbf{H}_2 \cdot \mathbf{M}_1 - \mathbf{E}_1 \cdot \mathbf{J}_2 + \mathbf{H}_1 \cdot \mathbf{M}_2 + \mathbf{E}_2 \cdot \mathbf{J}_1 \tag{12.1.11}$$

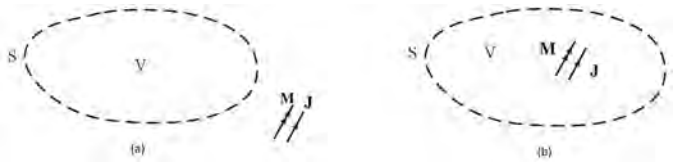


Figure 12.3: The geometry for proving reciprocity theorem when the surface S: (a) does not enclose the sources, and (b) encloses the sources. In the figure, the sources are supposed to be either $(\mathbf{M}_1, \mathbf{J}_1)$ producing fields $(\mathbf{E}_1, \mathbf{H}_1)$ or $(\mathbf{M}_2, \mathbf{J}_2)$ producing fields $(\mathbf{E}_2, \mathbf{H}_2)$.

[3]It is to be noted that in matrix algebra, the dot product between two vectors are often written as $\mathbf{a}^t \cdot \mathbf{b}$, but in the physics literature, the transpose on $\mathbf{a}$ is implied. Therefore, the dot product between two vectors is just written as $\mathbf{a} \cdot \mathbf{b}$.

12.1.1 Lorentz Reciprocity Theorem

First, we imagine a surface S that is bounding a volume V. In this sense, S is arbitrary as long as it excludes or encloses the sources as shown in (a) and (b) in Figure 12.3. Now, integrating (12.1.11) over the volume V, and invoking Gauss' divergence theorem, we have the reciprocity relationship that

$$\oiint_S d\mathbf{S} \cdot (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) = -\iiint_V dV[\mathbf{H}_2 \cdot \mathbf{M}_1 + \mathbf{E}_1 \cdot \mathbf{J}_2 - \mathbf{H}_1 \cdot \mathbf{M}_2 - \mathbf{E}_2 \cdot \mathbf{J}_1] \tag{12.1.12}$$

When the volume V contains no sources (see Figure 12.3), the reciprocity relatinoship reduces to

$$\oiint_S d\mathbf{S} \cdot (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) = 0 \tag{12.1.13}$$

The above is also called Lorentz reciprocity theorem by some authors.[4] It can be used to prove reciprocal relations between modes in a source-free waveguide, since no sources are involved in this reciprocity theorem.

12.1.2 Reaction Reciprocity Theorem

Next, when the surface S contains all the sources (see Figure 12.3), then the right-hand side of (12.1.12) will not be zero. On the other hand, when the surface $S \to \infty$, $\mathbf{E}_1$ and $\mathbf{H}_2$ becomes spherical waves which can be approximated by plane waves sharing the same $\boldsymbol{\beta}$ vector (also known as $\mathbf{k}$ vector or wave vector). Moreover, under the plane-wave approximation, in Maxwell's equations, we can replace ∇ with $-j\boldsymbol{\beta}$, and have $\omega\mu_0\mathbf{H}_2 = \boldsymbol{\beta} \times \mathbf{E}_2$, $\omega\mu_0\mathbf{H}_1 = \boldsymbol{\beta} \times \mathbf{E}_1$. Subsequently,

$$\mathbf{E}_1 \times \mathbf{H}_2 \sim \mathbf{E}_1 \times (\boldsymbol{\beta} \times \mathbf{E}_2) = \mathbf{E}_1(\boldsymbol{\beta} \cdot \mathbf{E}_2) - \boldsymbol{\beta}(\mathbf{E}_1 \cdot \mathbf{E}_2) \tag{12.1.14}$$

$$\mathbf{E}_2 \times \mathbf{H}_1 \sim \mathbf{E}_2 \times (\boldsymbol{\beta} \times \mathbf{E}_1) = \mathbf{E}_2(\boldsymbol{\beta} \cdot \mathbf{E}_1) - \boldsymbol{\beta}(\mathbf{E}_2 \cdot \mathbf{E}_1) \tag{12.1.15}$$

where the BAC-CAB formula has been used to simplify the above. But $\boldsymbol{\beta} \cdot \mathbf{E}_2 = \boldsymbol{\beta} \cdot \mathbf{E}_1 = 0$ in the far field since these far fields resemble plane waves. Furthermore, the $\boldsymbol{\beta}$ vectors are parallel to each other. Therefore, the two terms on the left-hand side of (12.1.12) cancel each other, and it vanishes when $S \to \infty$. (They cancel each other so that the remnant field vanishes faster than $1/r^2$. This is necessary[5] as the surface area S is growing larger and proportional to r^2.)

As a result, when $S \to \infty$, the left-hand side of (12.1.12) is zero, and it can be rewritten simply as

$$\int_V dV[\mathbf{E}_2 \cdot \mathbf{J}_1 - \mathbf{H}_2 \cdot \mathbf{M}_1] = \int_V dV[\mathbf{E}_1 \cdot \mathbf{J}_2 - \mathbf{H}_1 \cdot \mathbf{M}_2] \tag{12.1.16}$$

The inner product symbol is often used to rewrite the above as[6]

$$\langle \mathbf{E}_2, \mathbf{J}_1 \rangle - \langle \mathbf{H}_2, \mathbf{M}_1 \rangle = \langle \mathbf{E}_1, \mathbf{J}_2 \rangle - \langle \mathbf{H}_1, \mathbf{M}_2 \rangle \tag{12.1.17}$$

[4] Harrington, Time-Harmonic Electric Field [53].

[5] This is a mistake often committed by students of the course.

[6] Previously, we have used angular bracket to mean time average, but here, it means inner product.

where the inner product[7] $\langle \mathbf{A}, \mathbf{B} \rangle = \int_V dV \mathbf{A}(\mathbf{r}) \cdot \mathbf{B}(\mathbf{r})$.

Thus, the above inner product is also called **reaction** or the **reaction inner product**, a concept introduced by Rumsey [100]. The above is also called the **Rumsey reaction theorem**. Sometimes, the above is rewritten more succinctly or tersely as

$$\langle 2, 1 \rangle = \langle 1, 2 \rangle \tag{12.1.18}$$

where

$$\langle 2, 1 \rangle = \langle \mathbf{E}_2, \mathbf{J}_1 \rangle - \langle \mathbf{H}_2, \mathbf{M}_1 \rangle \tag{12.1.19}$$

The concept of inner product or reaction can be thought of as a kind of "measurement". The reciprocity theorem can be stated as that the fields generated by sources 2 as "measured" by sources 1 is equal to fields generated by sources 1 as "measured" by sources 2. This measurement concept is more lucid if we think of these sources as point sources or as Hertzian dipoles.

12.2 Conditions for Reciprocity

It is seen that the above proof hinges on (12.1.10) where $\overline{\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}^t$ and $\overline{\boldsymbol{\varepsilon}} = \overline{\boldsymbol{\varepsilon}}^t$. In other words, the anisotropic medium has to be described by symmetric tensors. They include conductive media, but not gyrotropic media that we studied previously; they are non-reciprocal. A ferrite biased by a magnetic field is often used in electronic circuits, and it corresponds to a gyrotropic, non-reciprocal medium.[8] Also, our starting equations (12.1.1) to (12.1.4) assume that the medium and the equations are linear time invariant so that Maxwell's equations can be written down in the frequency domain easily. Hence, the important conditions for reciprocity for a linear medium are

$$\overline{\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}^t, \qquad \overline{\boldsymbol{\varepsilon}} = \overline{\boldsymbol{\varepsilon}}^t \tag{12.2.1}$$

They include lossy anisotropic conductive media. Moreover, the medium is assumed stationary so that it is time-invariant.

[7] This inner product is quite different from those defined by mathematicians and physicists. We shall call this inner product the reaction inner product.

[8] Non-reciprocal media are important for making isolators in microwave. In an isolator, microwave signals can travel from Port 1 to Port 2, but not vice versa.

12.3 Application to a Two-Port Network and Circuit Theory

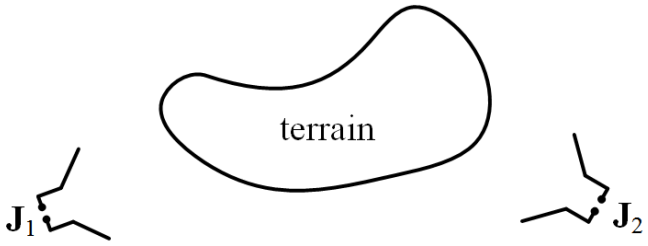


Figure 12.4: A geometry for proving the circuit relationship between two antennas using reciprocity theorem. Circuit relationship is possible when the ports of the antennas are small compared to wavelength. For these ports then, circuit theory prevails.

Figure 12.5: The terrain forming the media between the two antenna ports in Figure 12.4 can be as complicated as one can imagine, as long as the medium is reciprocal, and stationary (time-invariant). The one on the left shows a beautiful terrain with a whiff of civilization at the top of the hill, while the one on the right is a complicated, man-made printed circuit board with conductive elements. Circuit relationship is possible when the ports of the antennas are small compared to wavelength (courtesy of Marek Piwnicki and microcontrollerlab.com).

The reciprocity theorem can be used to distill and condense the interaction and relationship between two antennas over a complex terrain as long as the terrain comprises reciprocal media, namely, if $\overline{\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}^t$ and $\overline{\boldsymbol{\varepsilon}} = \overline{\boldsymbol{\varepsilon}}^t$ for these media.[9] In Figure 12.4, we assume that antenna 1 is driven by impressed current $\mathbf{J}_1$ while antenna 2 is driven by impressed current $\mathbf{J}_2$. It is assumed that the supports of these impressed currents are very small compared to wavelength so that circuit theory prevails at the antenna ports.[10] Further, it is assumed that the antennas are made from reciprocal media, such as conductive media. Since the system is linear time invariant, it can be written as the

[9] It is to be noted that a gyrotropic medium considred in Section 9.1 does not safisfy this reciprocity criteria, but it does satisfy the lossless criteria of Section 10.3.2.

[10] It can be shown that when the frequency is low, or the wavelength is long, then one can replace electromagnetic theory with electro-quasistatic theory or magneto-quasistatic theory. Or in short, by circuit theory.

interaction between two ports as in circuit theory as shown in Figure 12.6. Since these two ports are small compared to wavelengths, in the neighborhood of the ports, then we can apply circuit concepts like potential theory by letting $\mathbf{E} = -\nabla\Phi$ (see (3.3.7)). Thus, we can define voltages and currents at these ports, and V-I relationships can be established in the manner of circuit theory.

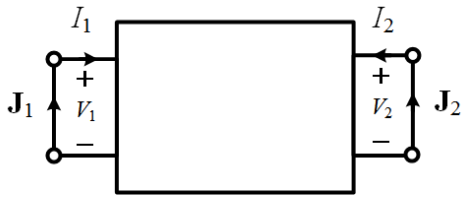


Figure 12.6: The interaction between two antennas in the far field of each other can be reduced to a circuit theory description since the input and output ports of the antennas are small compared to wavelength. But inside the box, any linear time invariant medium can be there.

Focusing on a two-port network as shown in Figure 12.6, circuit theory implies that [101]

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix} \tag{12.3.1}$$

This form is permissible since we have a linear time-invariant system, and this is the most general way to establish a linear relationship between the voltages and the currents at the two ports. This is quite obvious in a network of circuit elements, but this remains true for a general medium is harder to fathom: But it can be proved from electromagnetic theory. Furthermore, the matrix elements Z_{ij} can be obtained by performing a series of open-circuit and short-circuit measurements as in circuit theory.

Then assuming that the port 2 is turned on with $\mathbf{J}_2 \neq 0$, while port 1 is turned off with $\mathbf{J}_1 = 0$. In other words, port 1 is open circuit, and the source $\mathbf{J}_2$ is an impressed current source[11] that will produce an electric field $\mathbf{E}_2$ at port 1. Since the current at port 1 is turned off, or that $\mathbf{J}_1 = 0$, the voltage measured at port 1 is the open-circuit voltage V_1^{oc}. Please note here that $\mathbf{J}_1$ and $\mathbf{J}_2$ are impressed currents and are only defined in their respective ports. Moreover, in the long-wavelength limit, the currents are constant in the wires that carry them. Consequently, the reaction between $\mathbf{E}_2$ and $\mathbf{J}_1$ is

$$\langle \mathbf{E}_2, \mathbf{J}_1 \rangle = \int_V dV(\mathbf{E}_2 \cdot \mathbf{J}_1) = I_1 \int_{\text{Port 1}} \mathbf{E}_2 \cdot d\mathbf{l} = -I_1 V_1^{oc} \tag{12.3.2}$$

In proving reciprocity, we perform two Gedanken experiments consecutively. Even though port 1 is assumed to be off with no current through it, the $\mathbf{J}_1$ above is the impressed current to be used when port 1 is turned on. You need to contemplate on this a bit to wrap your head around this point!

[11] This is the same as the current source concept in circuit theory.

Since we assume the currents in the wire to be constant, then the current $\mathbf{J}_1$ is a constant current at the port when it is turned on. Or the current I_1 can be taken outside the integral. In slightly more details, the current $\mathbf{J}_1 = \hat{l} I_1/A$ where A is the cross-sectional area of the wire, and $\hat{l}$ is a unit vector aligned with the axis of the wire. The volume integral $dV = Adl$, and hence the second equality follows in the derivation above, where $d\mathbf{l} = \hat{l}dl$. Since $\int_{\text{Port 1}} \mathbf{E}_2 \cdot d\mathbf{l} = -V_1^{oc}$, we arrive at the last equality above.

We can repeat the derivation with port 2 to arrive at the reaction

$$\langle \mathbf{E}_1, \mathbf{J}_2 \rangle = I_2 \int_{\text{Port 2}} \mathbf{E}_1 \cdot d\mathbf{l} = -I_2 V_2^{oc} \tag{12.3.3}$$

Reciprocity requires these two reactions to be equal, and therefore,

$$I_1 V_1^{oc} = I_2 V_2^{oc} \tag{12.3.4}$$

But from (12.3.1), we can set the pertinent currents to zero to find these open circuit voltages to be used in (12.3.4). Therefore, $V_1^{oc} = Z_{12} I_2$, $V_2^{oc} = Z_{21} I_1$. Since $I_1 V_1^{oc} = I_2 V_2^{oc}$ by the reaction concept or by reciprocity, then $Z_{12} = Z_{21}$. The above analysis can be easily generalized to an N-port network.

The simplicity of the above belies its importance. The above shows that the reciprocity concept in circuit theory is a special case of reciprocity theorem for electromagnetic theory. The terrain can also be replaced by complex circuits as in a circuit board, as long as the materials in the terrain or circuit board are reciprocal, linear and time invariant. For instance, the complex terrain can also be replaced by complex antenna structures. It is to be noted that even when the transmit and receive antennas are miles apart, as long as the transmit and receive ports of the linear time invariant system can be characterized by a linear relation expounded by (12.3.1), and the ports small enough compared to wavelength so that circuit theory prevails at the ports, we can apply the above analysis! This relation that $Z_{12} = Z_{21}$ is true as long as the medium traversed by the fields is a reciprocal medium even though the ports may be far apart.

Before we conclude this section, it is to be mentioned that some researchers advocate the use of circuit theory to describe electromagnetic theory. Such is the case in the transmission line matrix (TLM) method [102], and the partial element equivalence circuit (PEEC) method [103]. Circuit theory is so simple that many people fall in love with it!

12.4 Voltage Sources in Electromagnetics

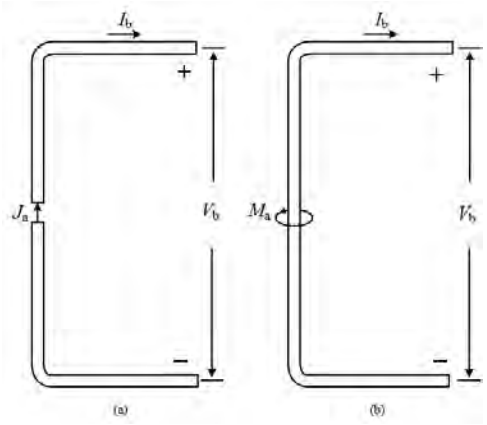


Figure 12.7: Two ways to model electromagnetic sources: (i) A current source modeled by an impressed current source $\mathbf{J}_a$ driving a very short antenna, and (ii) A voltage source modeled by an impressed magnetic frill source (loop source) $\mathbf{M}_a$ driving a very short antenna (courtesy of Kong, Electromagnetic Wave Theory [34]).

In the above discussions, we have used impressed current sources in reciprocity theorem to derive certain circuit concepts. (As mentioned before, impressed currents are immutable with respect to change in its environment, just as current sources are immutable in circuit theory.)

Before we end this section, it is prudent to mention how voltage sources are modeled in electromagnetic theory. They have to be immutable with respect to change in its environment, and can be modeled by impressed magnetic currents.

The use of the impressed currents so that circuit concepts can be applied is shown in Figure 12.7. The antenna in (a) is driven by a current source. But a magnetic current source (loop) can be used as a voltage source in circuit theory as shown by Figure 12.7b. By using duality concept, Faraday's law with magnetic current is similar to Ampere's law with electric current. An electric field has to curl around a magnetic current in Faraday's law; dual to the case of Ampere's law where magnetic field curls around an electric current. This electric field will cause a voltage drop between the metal above and below the magnetic current loop making it behave like a voltage source.[12]

12.5 Hind Sight

The proof of reciprocity theorem for Maxwell's equations is very deeply related to the symmetry of the operator involved. We can elucidate this from linear algebra. Given a linear system which

[12]More discussions can be found in Jordain and Balmain, Electromagnetic Waves and Radiation Systems [57].

can be modeled by a matrix equation driven by two right-hand sides, $\mathbf{b}_1$ and $\mathbf{b}_2$ with solutions $\mathbf{x}_1$ and $\mathbf{x}_2$, they can be written succinctly as

$$\overline{\mathbf{A}} \cdot \mathbf{x}_1 = \mathbf{b}_1 \tag{12.5.1}$$

$$\overline{\mathbf{A}} \cdot \mathbf{x}_2 = \mathbf{b}_2 \tag{12.5.2}$$

We can left dot multiply the first equation with $\mathbf{x}_2$ and do the same with the second equation with $\mathbf{x}_1$ to arrive at

$$\mathbf{x}_2^t \cdot \overline{\mathbf{A}} \cdot \mathbf{x}_1 = \mathbf{x}_2^t \cdot \mathbf{b}_1 \tag{12.5.3}$$

$$\mathbf{x}_1^t \cdot \overline{\mathbf{A}} \cdot \mathbf{x}_2 = \mathbf{x}_1^t \cdot \mathbf{b}_2 \tag{12.5.4}$$

If $\overline{\mathbf{A}}$ is symmetric, the left-hand side of both equations are equal to each other.[13] Therefore, we can equate their right-hand sides to arrive at

$$\mathbf{x}_2^t \cdot \mathbf{b}_1 = \mathbf{x}_1^t \cdot \mathbf{b}_2 \tag{12.5.5}$$

The above is analogous to the statement of the reciprocity theorem which is

$$\langle \mathbf{E}_2, \mathbf{J}_1 \rangle = \langle \mathbf{E}_1, \mathbf{J}_2 \rangle \tag{12.5.6}$$

where the reaction inner product is $\langle \mathbf{E}_i, \mathbf{J}_j \rangle = \int_V dV \mathbf{E}_i(\mathbf{r}) \cdot \mathbf{J}_j(\mathbf{r})$ as mentioned before,. The inner product in linear algebra is that of dot product in matrix theory, but the inner product for reciprocity theorem is that for infinite dimensional spaces.[14] So if the operators in Maxwell's equations are symmetrical, then reciprocity theorem applies.

It is prudent to mention that in linear algebra, for two vectors $\mathbf{a}$ and $\mathbf{b}$, there are two kinds of dot products or inner products. They are written as

$$\mathbf{a}^t \cdot \mathbf{b} = \sum_i a_i b_i$$

and

$$\mathbf{a}^\dagger \cdot \mathbf{b} = \sum_i a_i^* b_i$$

When $\mathbf{a} = \mathbf{b}$, the second inner product ensures that the it is positive definite. We shall call the second inner product energy inner product, while the first inner product, the reaction inner product used in reciprocity.

In the infinite dimensional continuum space (or Hilbert space), analogously, the above inner products between vector functions $\mathbf{a}(\mathbf{r})$ and $\mathbf{b}(\mathbf{r})$ are

$$\langle \mathbf{a}(\mathbf{r}), \mathbf{b}(\mathbf{r}) \rangle = \int dV \mathbf{a}(\mathbf{r}) \cdot \mathbf{b}(\mathbf{r})$$

and

$$\langle \mathbf{a}^*(\mathbf{r}), \mathbf{b}(\mathbf{r}) \rangle = \int dV \mathbf{a}^*(\mathbf{r}) \cdot \mathbf{b}(\mathbf{r})$$

In many math literature, the complex conjugation is implied in the inner products, but not in electromagnetics literature. Notice that now, $\mathbf{r}$ replaces the role of i in the discrete case.

[13] This can be easily proven by taking the transpose of a scalar, and taking the transpose of the product of matrices.

[14] Such spaces are called Hilbert space.

Exercises for Lecture 12

Problem 12-1: Show that the left-hand side of (12.1.12) is in fact zero by showing that (12.1.14) and (12.1.15) are true in the far field.

Problem 12-2:

(i) Explain the difference in reciprocity expressed via the Lorentz reciprocity theorem and the Rumsey reaction theorem.

(ii) Show that in a reciprocity linear circuit containing N ports which can be modeled by an $N \times N$ impedance matrix $\overline{\mathbf{Z}}$, reciprocity means that $Z_{ij} = Z_{ji}$ where Z_{ij} is the ij-element of the matrix $\overline{\mathbf{Z}}$.

(iii) Show that the reciprocity theorem is related to a symmetric matrix system in linear algebra.

Problem 12-3: Explain why a magnetic loop around a metal rod can be used to model a voltage source.Faf

L

Chapter 13

Equivalence Theorems, Huygens' Principle

Electromagnetic equivalence theorems are useful for simplifying solutions to many problems. The rule of physics and engineering is that if we encounter a problem that is hard to fathom, we distill it down to a combination of sum of smaller problems that are easier to solve. In this manner, highly complex problems can be solved. With the help of computers, highly complicated problems can be solved. Also, they offer physical insight into the behaviour of electromagnetic fields of a Maxwellian system. They are closely related to Huygens' principle. One application is their use in studying the radiation from an aperture antenna or from the output of a lasing cavity. These theorems are discussed in many textbooks [53, 57, 68, 34, 104]. Some authors also call them Love's equivalence principles [105] and credit has been given to Schelkunoff as well [93].

You may have heard of another equivalence theorem in special relativity. It was postulated by Einstein to explain why light ray should bend around a star. The equivalence theorem in special relativity is very different from that in electromagnetics.

13.1 Equivalence Theorems or Equivalence Principles

13.1.1 Inside-Out Case

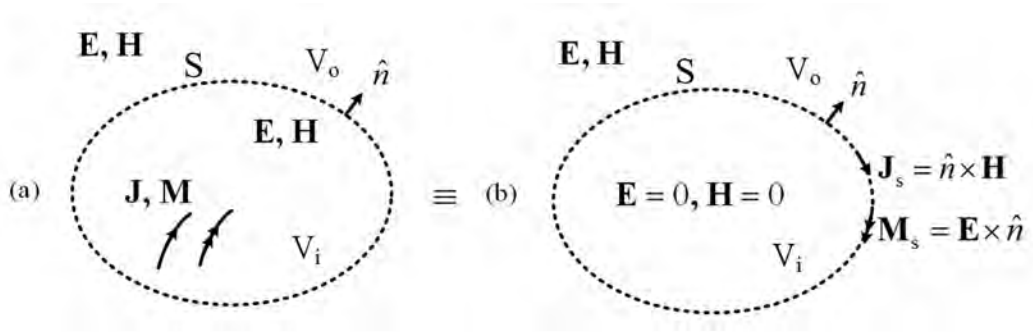


Figure 13.1: The inside-out problem where the two cases in (a) and (b) are equivalent. In (b), equivalence currents are impressed on the surface S so as to produce the same fields outside in V_o in both cases (a) and (b).

In this lecture, we will consider three equivalence theorems: (1) The inside out case. (2) The outside in case. (3) The general case. As mentioned above, we will derive these theorems using thought experiments or Gedanken experiments. As shall be shown later, they can also be derived mathematically using Green's theorem.

We will examine case (1) first. In this case, as shown in Figure 13.1(a), we let $\mathbf{J}$ and $\mathbf{M}$ be the time-harmonic radiating sources inside a surface S radiating into a region $V = V_o \cup V_i$.[1] They produce $\mathbf{E}$ and $\mathbf{H}$ everywhere. These fields $\mathbf{E}$ and $\mathbf{H}$ are unique provided that the Sommerfeld radiation condition is satisfied.

We call these fields and sources Maxwellian since they satisfy Maxwell's equations. We postulate an equivalence problem indicated in Figure 13.1(b) by first constructing an imaginary surface S. In this equivalence problem, the fields outside S in V_o are the same in both (a) and (b). But in (b), the fields inside S in V_i are zero. Despite, the fields and sources in (b) are Maxwellian to be explained next.

To explain further, in Figure 13.1(b), the tangential components of the fields are discontinuous at S. This is not possible for a Maxwellian fields unless surface currents are impressed on the surface S. We have learned from electromagnetic boundary conditions that electromagnetic fields are discontinuous across a current sheet. Then we ask ourselves what surface currents are needed on surface S so that the boundary conditions (or jump condition) for field discontinuities are satisfied on S. Clearly, surface currents are needed for these field discontinuities. Since both the $\mathbf{E}$ and the $\mathbf{H}$ fields are discontinuous, we need impressed $\mathbf{J}_s$ and $\mathbf{M}_s$ at the interface to account for the jump discontinuities. This will make the fields and sources in (b) Maxwellian, i.e., they are also solutions to Maxwell's equations driven by different sources.

[1] This is the math notation for "union", the parlance for "sum".

By virtue of the boundary conditions and the jump conditions in electromagnetics, these surface currents to be impressed on S are

$$\mathbf{J}_s = \hat{n} \times \mathbf{H}, \qquad \mathbf{M}_s = \mathbf{E} \times \hat{n} \tag{13.1.1}$$

We have learnt from Section 4.3.3 that an electric current sheet in Ampere's law produced a jump discontinuity in the magnetic field. By the same token, fictitious magnetic current is added to Faraday's law in Section 5.3 for mathematical symmetry. Then by duality, a magnetic current sheet induces a jump discontinuity in the electric field. Because of the opposite polarity of the magnetic current $\mathbf{M}$ in Faraday's law compared to the electric current $\mathbf{J}$ in Ampere's law as is shown in Section 5.3, this magnetic current sheet is proportional to $\mathbf{E} \times \hat{n}$ instead of $\hat{n} \times \mathbf{H}$.

Consequently, we can convince ourselves that $\hat{n} \times \mathbf{H}$ and $\mathbf{E} \times \hat{n}$ just outside S in both cases (a) and (b) are the same. Furthermore, we are persuaded that the above is a bona fide solution to Maxwell's equations. The case (a) in Figure 13.1 satisfies Maxwell's equations with current sources $\mathbf{J}$ and $\mathbf{M}$ inside V_i and for $\mathbf{E}$ and $\mathbf{H}$ everywhere.

In case (b), the $\mathbf{E}$ and $\mathbf{H}$ fields satisfy Maxwell's equations with the impressed surface current sources $\mathbf{J}_s$ and $\mathbf{M}_s$ on S, but with the original sources $\mathbf{J}$ and $\mathbf{M}$ removed in V_i. In case (b), the fields inside V_i is set to zero but $\mathbf{E}$ and $\mathbf{H}$ are the same as case (a) outside (or in V_0). Thus, there are discontinuous fields on the surface S, but the discontinuities are supported by the impressed currents $\mathbf{J}_s$ and $\mathbf{M}_s$. Hence, the fields and sources in (b) satisfy Maxwell's equations, or that they are Maxwellian.

Next, we have to convince ourselves that the fields outside S are the same in both cases. This follows from the uniqueness theorem: the fields in both cases satisfy the same boundary conditions and the radiation condition at infinity. It seems that there are some redundancy here, since both the boundary conditions for $\mathbf{E}$ and $\mathbf{H}$ are met here. But as long as these fields are consistent, that is okay. These fields are consistent since they are Maxwellian.

The above fact can be proved mathematically, as shall be shown later by a more mathematical manipulation. The fact that the fields are zero in V_i or inside S is known as the *extinction theorem*. This equivalence theorem and extinction theorem can also be proved mathematically.

13.1.2 Outside-in Case

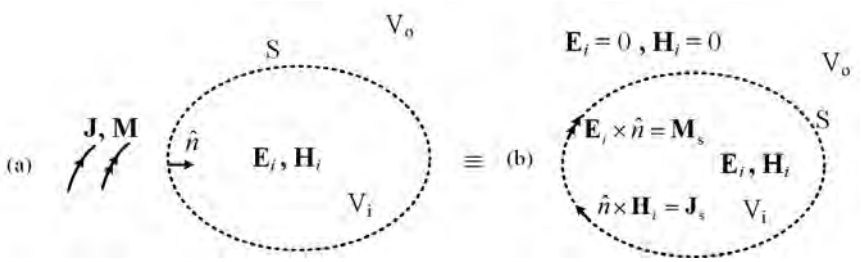


Figure 13.2: The outside-in problem where equivalence currents are impressed on the surface S to produce the same fields inside in both cases.

Similar to before, in Figure 13.2, we find an equivalence problem (b) where the fields inside S in V_i are the same as in (a), but the fields outside S in V_o in the equivalence problem is zero. The fields are discontinuous across the surface S, and hence, impressed surface currents $\mathbf{J}_s$ and $\mathbf{M}_s$ are needed to account for these discontinuities.

Then by the uniqueness theorem,[2] the fields $\mathbf{E}_i$, $\mathbf{H}_i$ inside V in both cases are the same. Again, by the *extinction theorem*, the fields produced by $\mathbf{E}_i \times \hat{n}$ and $\hat{n} \times \mathbf{H}_i$ are zero outside S.

It is to be noted that for both inside-out and outside-in cases, the field is extinct by the extinction theorem only in the volume or region that originally contains the sources. This will be clear when these equivalence problems are derived mathematically.

13.1.3 General Case

From these two cases, we can create a rich variety of equivalence problems. By linear superposition of the inside-out problem, and the outside-in problem, then a third equivalence problem is shown in Figure 13.3:

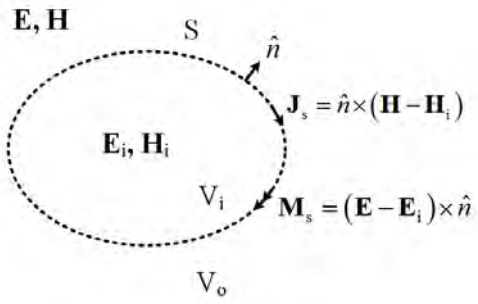


Figure 13.3: The general case where the fields are non-zero both inside and outside the surface S. Equivalence currents are needed on the surface S to support the jump discontinuities in the fields.

13.2 Electric Current on a PEC–Relation to Uniqueness Theorem

In this section, we show that new equivalence problems can be derived if the equivalence surface currents are radiating on the surface of a PEC. This can be obtained by using the equivalence problems in the previous section, to derive other corollaries of equivalence problems. We also need a result from reciprocity theorem, that impressed surface currents on the PEC cannot radiate.

We can start with the inside-out equivalence problem as shown in (b) of Figure 13.1. Since the fields inside S is zero for the inside-out problem, using a Gedanken experiment , one can insert a PEC object inside S without disturbing the fields $\mathbf{E}$ and $\mathbf{H}$ outside since the field is zero inside S.

[2] We can add infinitesimal loss to ensure that uniqueness theorem is satisfied in this enclosed volume V_i.

As the PEC object grows to snugly fit inside the surface S, then the electric current $\mathbf{J}_s = \hat{n} \times \mathbf{H}$, which is an impressed current source on top of a PEC, does not radiate by reciprocity theorem. Only one of the two impressed currents is radiating, namely, the magnetic current $\mathbf{M}_s = \mathbf{E} \times \hat{n}$ is radiating; and hence, $\mathbf{J}_s$ in Figure 13.4 can be removed. This is commensurate with the uniqueness theorem that only the knowledge of $\mathbf{E} \times \hat{n}$ plus the radiation condition, are needed to uniquely determine the fields outside S.

It is to be noted that $\mathbf{J}_s$, $\mathbf{M}_s$, $\mathbf{E}$ and $\mathbf{H}$ form a Maxwellian system before we insert a PEC object inside the surface S shown in (b) in Figure 13.1.

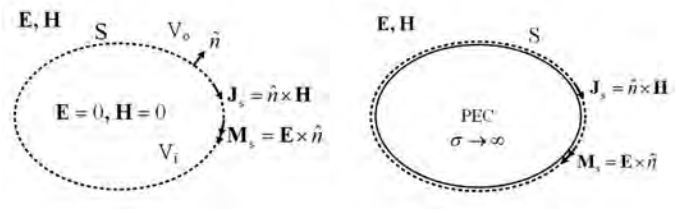


Figure 13.4: For the figure on the left, since the field is zero inside S, a PEC can be inserted inside S without disturbing the field outside as shown on the right. But the electric current does not radiate on a PEC surface. Thus only one of the two currents is needed.

13.3 Magnetic Current on a PMC–Relation to Uniqueness Theorem

Again, from reciprocity, an impressed magnetic current on a PMC cannot radiate. By the same token, we can perform the Gedanken experiment as before by inserting a PMC object inside S. It will not alter the fields outside S, as the fields inside S is zero. As the PMC object grows to snugly fit the surface S, only the impressed electric current $\mathbf{J}_s = \hat{n} \times \mathbf{H}$ radiates, and the impressed magnetic current $\mathbf{M}_s = \mathbf{E} \times \hat{n}$ does not radiate and it can be removed. This is again commensurate with the uniqueness theorem that only the knowledge of the $\hat{n} \times \mathbf{H}$ is needed to uniquely determine the fields outside S.

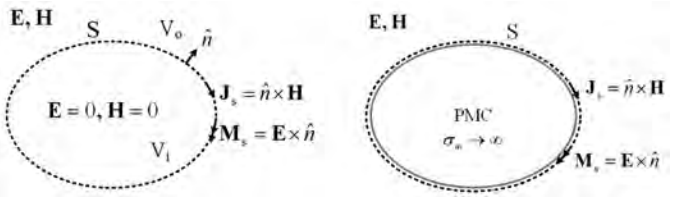


Figure 13.5: Similarly, a Gedanken experiment can be done by inserting a PMC inside S. By expanding the PMC surface to snugly fit inside S, only an electric current is needed to produce the field outside the surface S.

13.4 Huygens' Principle and Green's Theorem

Huygens' principle shows how a wave field on a surface determines the wave field outside the surface S. This concept was expressed by Huygens heuristically in the 1600s [106]. But the mathematically precise expression of this idea was due to George Green[3] in the 1800s. This concept can be expressed mathematically for both scalar and vector waves. The derivation for the vector wave case is "homomorphic" to the scalar wave case. But the algebra in the scalar wave case is much simpler. Therefore, we shall study first the scalar wave case first with simpler algebra, followed by the electromagnetic vector wave case later where the vector algebra is more complex.

[3]George Green (1793-1841) was self educated and the son of a miller in Nottingham, England [107]. If you visit Nottingham, you will see a windmill built in his honor.

13.4.1 Scalar Waves Case

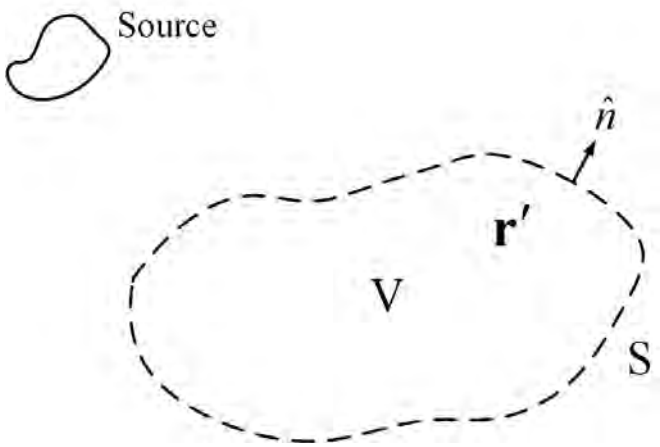


Figure 13.6: The geometry for deriving Huygens' principle for scalar wave equation. The dashed surface represents an "imagined" surface.

For a $\Psi(\mathbf{r})$ that satisfies the scalar wave equation inside V surrounded by an "imagined" surface S, then

$$(\nabla^2 + k^2)\,\Psi(\mathbf{r}) = 0, \qquad \mathbf{r} \in V \tag{13.4.1}$$

Notice that V does not contain the source that produces $\Psi(\mathbf{r})$ so that the right-hand side of (13.4.1) can be set to zero always where $\mathbf{r} \in V$. The corresponding scalar Green's function $g(\mathbf{r}, \mathbf{r}')$ for a homogeneous medium satisfies

$$(\nabla^2 + k^2)\,g(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}') \quad \forall \mathbf{r}, \quad \forall \mathbf{r}'. \tag{13.4.2}$$

which is the point source response in a homogeneous medium.[4] Next, we multiply (13.4.1) by $g(\mathbf{r}, \mathbf{r}')$ and (13.4.2) by $\Psi(\mathbf{r})$. Upon substracting them, we have

$$\Psi(\mathbf{r})\delta(\mathbf{r} - \mathbf{r}') = g(\mathbf{r}, \mathbf{r}')\nabla^2\Psi(\mathbf{r}) - \Psi(\mathbf{r})\nabla^2 g(\mathbf{r}, \mathbf{r}') \tag{13.4.3}$$

Notice that the term involving $gk^2\Psi$ cancel each other. As shall be shown, the right-hand side can be written as the divergence of a vector field, motivating the use of Gauss' divergence theorem. To this end, we integrate resultant equation and over a volume V as shown in Figure 13.6.

There are two cases to consider: when (1) $\mathbf{r}'$ is in V, or (2) when $\mathbf{r}'$ is outside V. Thus, we have the dichotomous result on the left-hand side of the equation, viz.,

$$\left.\begin{array}{ll} \text{if } \mathbf{r}' \in V, & \Psi(\mathbf{r}') \\ \text{if } \mathbf{r}' \notin V, & 0 \end{array}\right\} = \int_V d\mathbf{r}\,[g(\mathbf{r}, \mathbf{r}')\nabla^2\Psi(\mathbf{r}) - \Psi(\mathbf{r})\nabla^2 g(\mathbf{r}, \mathbf{r}')], \tag{13.4.4}$$

[4] It can be shown that in a homogeneous medium, this scalar Green's function is given by $g(\mathbf{r}, \mathbf{r}') = 1/(4\pi|\mathbf{r} - \mathbf{r}'|)\exp(-jk|\mathbf{r} - \mathbf{r}'|)$. [108][p. 24-26]

The left-hand side evaluates to different values depending on where $\mathbf{r}'$ is due to the sifting property of the delta function $\delta(\mathbf{r}-\mathbf{r}')$. Since $g\nabla^2\Psi - \Psi\nabla^2 g = \nabla\cdot(g\nabla\Psi - \Psi\nabla g)$, the right-hand side of (13.4.4) can be rewritten using Gauss' divergence theorem, giving

$$\left.\begin{array}{ll} \text{if } \mathbf{r}'\in V, & \Psi(\mathbf{r}') \\ \text{if } \mathbf{r}'\notin V, & 0 \end{array}\right\} = \oint_S dS\,\hat{n}\cdot[g(\mathbf{r},\mathbf{r}')\nabla\Psi(\mathbf{r}) - \Psi(\mathbf{r})\nabla g(\mathbf{r},\mathbf{r}')], \tag{13.4.5}$$

where S is the surface bounding V. The above is the Green's theorem, or the mathematical expression that once $\Psi(\mathbf{r})$ and $\hat{n}\cdot\nabla\Psi(\mathbf{r})$ are known on S, then $\Psi(\mathbf{r}')$ away from S could be found. This is similar to the expression of equivalence principle where $\hat{n}\cdot\nabla\Psi(\mathbf{r})$ and $\Psi(\mathbf{r})$ are equivalence sources on the surface S. They can be used to find the fields inside and outside V, and the extinction theorem is beautifully embodied in this equation also as shall be shown.

The first term on the right-hand side radiates via the Green's function $g(\mathbf{r},\mathbf{r}')$ which radiates like a monopole source producing a spherically symmetric field. Since this is a monopole field, this source is also called a monolayer or single layer source. The second term radiates, on the other hand, via the normal derivative of the Green's function, namely $\hat{n}\cdot\nabla g(\mathbf{r},\mathbf{r}')$. Since the derivative of a Green's function yields a dipole field (see Problem 4-2), with the dipole pointing normal to the surface S, the second term corresponds to sources that radiate like dipoles. These sources are also called double layer (or dipole layer) sources. (These terminologies are prevalent in the acoustics and mathematics communities.) As aforementioned, the above mathematical expression also embodies the *extinction theorem* that says if $\mathbf{r}'$ is outside V, or $\mathbf{r}'\notin V$, the left-hand side evaluates to zero. In this case, the monolayer source and the dipole layer source in (13.4.5) produce fields that cancel each other outside V.

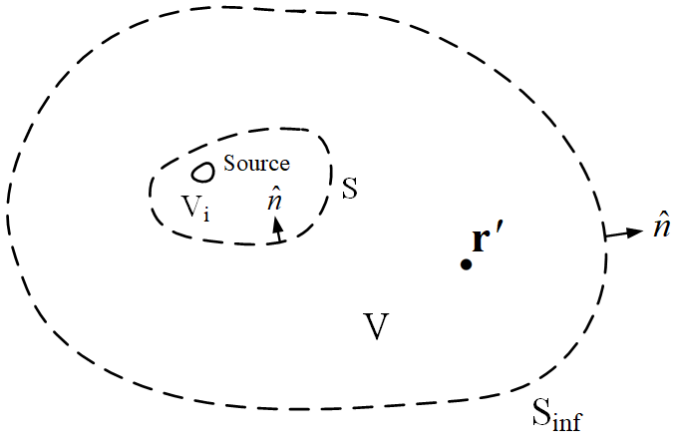


Figure 13.7: The geometry for deriving Huygens' principle for scalar wave. The radiation from the source can be replaced by equivalence sources on the surfaces S and S_{inf}, and the field inside V bounded by S and S_{inf} can be calculated using (13.4.5). Also, the field is zero inside S from (13.4.5). This is the extinction theorem.

Let us assume that the volume V is bounded by two "imagined" surfaces S and S_{inf} (with

$S_{\text{inf}} \to \infty$ eventually) as shown in Figure 13.7. Initially, the surface integral in (13.4.5) should include an integral over S_{inf}. But the integral over S_{inf} eventually vanishes as we shall explain next.

Remember that when $S_{\text{inf}} \to \infty$, all fields look like spherical waves which can be approximated by plane waves in the far field. Also, on the right-hand side of (13.4.5), the fields, $\Psi(\mathbf{r})$ on the surface S_{inf} in the integrand of (13.4.5) become plane waves, and $\nabla \to -\hat{r}jk$ on S_{inf}. Furthermore, for $g(\mathbf{r}-\mathbf{r}')$, $\mathbf{r}'$ remains finite, while $\mathbf{r}$ goes to infinity. Hence, $g(\mathbf{r}-\mathbf{r}') \sim O(1/r)$, when $r \to \infty$, which is a spherical wave morphing into a plane wave. Since $\Psi(\mathbf{r})$ is produced by finite sources in V_i, $\Psi(\mathbf{r}) \sim O(1/r)$, when $r \to \infty$.[5] It is also a spherical wave morphing into a plane wave, or $\nabla g \to -\hat{r}jkg$. Then, with some algebra, the integrand of the two terms on the right-hand side of (13.4.5) cancel each other on S_{inf}. With care, it can be shown that the left-over terms decay faster than $(1/r^2)$ integral over S_{inf} in (13.4.5) and eventually, vanishes.[6]

Therefore, the integral over S_{inf} becomes zero, and can be ignored, and (13.4.5) is valid for the case shown in Figure 13.7 as well but with the surface integral over surface S only. Physically, it implies that we are integrating over equivalence sources on S that radiate to infinity.

Here, the field outside S at $\mathbf{r}'$ is expressible in terms of the field on S. Therefore, this is similar to the inside-out equivalence principle we have discussed previously Section 13.1.1, albeit this is for scalar wave case.

Notice that in deriving (13.4.5), $g(\mathbf{r},\mathbf{r}')$ has only to satisfy (13.4.2) for both $\mathbf{r}$ and $\mathbf{r}'$ in V but no boundary condition has yet been imposed on $g(\mathbf{r},\mathbf{r}')$. Therefore, if we further require that $g(\mathbf{r},\mathbf{r}')=0$ for $\mathbf{r} \in S$, when $\mathbf{r}$ is in V. Then (13.4.5) becomes

$$\Psi(\mathbf{r}') = -\oint_S dS\, \Psi(\mathbf{r})\, \hat{n} \cdot \nabla g(\mathbf{r},\mathbf{r}'), \qquad \mathbf{r}' \in V. \tag{13.4.6}$$

On the other hand, if require additionally that $g(\mathbf{r},\mathbf{r}')$ satisfies (13.4.2) with the boundary condition $\hat{n} \cdot \nabla g(\mathbf{r},\mathbf{r}')=0$ for $\mathbf{r} \in S$, then (13.4.5) becomes

$$\Psi(\mathbf{r}') = \oint_S dS\, g(\mathbf{r},\mathbf{r}')\, \hat{n} \cdot \nabla \Psi(\mathbf{r}), \qquad \mathbf{r}' \in V. \tag{13.4.7}$$

Equations (13.4.5), (13.4.6), and (13.4.7) are various forms of Huygens' principle, or equivalence principle for scalar waves (acoustic waves) depending on the definition of $g(\mathbf{r},\mathbf{r}')$. Equations (13.4.6) and (13.4.7) stipulate that only $\Psi(\mathbf{r})$ or $\hat{n} \cdot \nabla \Psi(\mathbf{r})$ need be known on the surface S in order to determine $\Psi(\mathbf{r}')$. The above are analogous to the PEC and PMC equivalence principle considered previously in Sections 13.2 and 13.3.

[5] The symbol "O" means "of the order" in the math community. The computer science community uses a somewhat different notation.

[6] This decay rate faster than $(1/r^2)$ is necessary since the surface S_{inf} grows as r^2.

13.4.2 Electromagnetic Waves Case

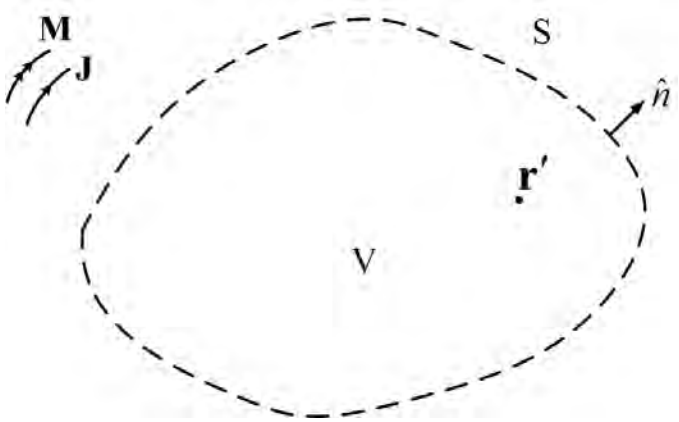


Figure 13.8: Derivation of the Huygens' principle for the electromagnetic case. One only needs to know the surface fields on surface S in order to determine the field at $\mathbf{r}'$ inside V.

The derivation of Huygens' principle and Green's theorem for the electromagnetic case is more complicated than the scalar wave case. But fortunately, this problem is mathematically "homomorphic" to the scalar wave case. In dealing with the requisite vector algebra, we have to remember to cross the t's and dot the i's, to carry ourselves carefully through the somewhat laborious and complicated vector algebra! One can always refer back to the scalar-wave case to keep our bearing straight.

In a source-free homogeneous region, an electromagnetic wave satisfies the vector wave equation

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r}) - k^2\, \mathbf{E}(\mathbf{r}) = 0, \qquad \mathbf{r} \in V. \tag{13.4.8}$$

The above also implies that the field $\mathbf{E}$ is Maxwellian. Again, we pick the volume V to contain no source so that the right-hand side of the above is zero when $\mathbf{r} \in V$. Comparing with the scalar wave case, the analogue of the scalar Green's function is the dyadic Green's function for the electromagnetic wave case [109, 1, 110, 34]. Moreover, the dyadic Green's function satisfies the equation[7]

$$\nabla \times \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') - k^2\, \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = \overline{\mathbf{I}}\, \delta(\mathbf{r} - \mathbf{r}'). \tag{13.4.9}$$

In the above, the source point is given by $\mathbf{r}'$ while the field point is given by $\mathbf{r}$. It can be shown by direct back substitution that the dyadic Green's function in free space is [110]

$$\overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = \left(\overline{\mathbf{I}} + \frac{\nabla\nabla}{k^2}\right) g(\mathbf{r}, \mathbf{r}') \tag{13.4.10}$$

[7] A dyad is an outer product between two vectors, and it behaves like a tensor, except that a tensor is more general than a dyad. A purist will call the above a tensor Green's function, as the above is not a dyad in its strictest definition.

The back substitution shows that as long as the scalar Green's function $g(\mathbf{r},\mathbf{r}')$ satisfies (13.4.2), the above is the solution to (13.4.9). The above allows us to derive the vector Green's theorem [109, 1, 34].[8] We will use the insight we have gained in deriving the scalar Green's theorem to derive the Huygens' principle for the vector field.

Then, after post-multiplying (13.4.8) by $\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')\cdot\mathbf{a}$, pre-multiplying[9] (13.4.9) by $\mathbf{E}(\mathbf{r})$, subtracting the resultant equations, the terms involving $k^2\mathbf{E}(\mathbf{r})\cdot\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')\cdot\mathbf{a}$ cancel out. We arrive at[10]

$$\mathbf{E}(\mathbf{r})\cdot\mathbf{a}\delta(\mathbf{r}-\mathbf{r}')=\mathbf{E}(\mathbf{r})\cdot\nabla\times\nabla\times\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')\cdot\mathbf{a}-\nabla\times\nabla\times\mathbf{E}(\mathbf{r})\cdot\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')\cdot\mathbf{a} \tag{13.4.11}$$

We then integrate the difference over volume V, and using the sifting property of the delta function, considering two cases as we did for the scalar wave case, we have

$$\left.\begin{array}{ll}\text{if } \mathbf{r}'\in V, & \mathbf{E}(\mathbf{r}')\\ \text{if } \mathbf{r}'\notin V, & 0\end{array}\right\}=\int_V dV\ \left[\mathbf{E}(\mathbf{r})\cdot\nabla\times\nabla\times\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')\right.$$
$$\left.-\nabla\times\nabla\times\mathbf{E}(\mathbf{r})\cdot\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')\right]. \tag{13.4.12}$$

Next, using the vector identity that[11]

$$-\nabla\cdot\left[\mathbf{E}(\mathbf{r})\times\nabla\times\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')+\nabla\times\mathbf{E}(\mathbf{r})\times\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')\right]$$
$$=\mathbf{E}(\mathbf{r})\cdot\nabla\times\nabla\times\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')-\nabla\times\nabla\times\mathbf{E}(\mathbf{r})\cdot\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}'), \tag{13.4.13}$$

then the integrand of (13.4.12) can be written as a divergence. With the help of Gauss' divergence theorem, the right-hand side of (13.4.12) can be written as

$$\left.\begin{array}{ll}\text{if } \mathbf{r}'\in V, & \mathbf{E}(\mathbf{r}')\\ \text{if } \mathbf{r}'\notin V, & 0\end{array}\right\}=-\oint_S dS\,\hat{n}\cdot\left[\mathbf{E}(\mathbf{r})\times\nabla\times\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')+\nabla\times\mathbf{E}(\mathbf{r})\times\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')\right]$$
$$=-\oint_S dS\,\left[-\mathbf{E}(\mathbf{r})\times\hat{n}\cdot\nabla\times\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')-j\omega\mu\,\hat{n}\times\mathbf{H}(\mathbf{r})\cdot\overline{\mathbf{G}}(\mathbf{r},\mathbf{r}')\right]. \tag{13.4.14}$$

Notice that to invoke the Gauss' divergence theorem, the bounding surface S above is one that includes the surface at infinity as shown in the geometry for the scalar wave case in Figure 13.7. Again, the dichotomous value on the left-hand side of the above follows from the sifting property of the $\delta(\mathbf{r}-\mathbf{r}')$. It is simple and embodies the extinction theorem. It would not have been that simple if the right-hand side of (13.4.8) has not been made zero with a proper choice of V. Hence, the field is extinct for $\mathbf{r}'\notin V$ or in the volume that contains the sources that generate the fields.

[8] To keep our sanity, it is prudent to convert (13.4.9) by an arbitrary vector $\mathbf{a}$ to convert it from a dyad equation to a vector equation so that we can understand them better. The arbitrary vector $\mathbf{a}$ can be cancelled after you have gone through the somewhat insane manipulations.

[9] Since we are dealing with dyads which are tensors like matrices, order is very important here.

[10] For those less mathematical inclined, you may want to take a small diversion to derive the following.

[11] This identity can be established by using the identity $\nabla\cdot(\mathbf{A}\times\mathbf{B})=\mathbf{B}\cdot\nabla\times\mathbf{A}-\mathbf{A}\cdot\nabla\times\mathbf{B}$. We will have to let (13.4.13) act on an aribitrary constant vector to convert the dyad into a vector before applying this identity. The equality of the volume integral in (13.4.12) to the surface integral in (13.4.14) is also known as the vector Green's theorem [109, 34]. Earlier form of this theorem was known as Franz formula [111].

The above is just the vector analogue of (13.4.5). We have used the cyclic relation of scalar-triple product to rewrite the last expression. Since $\hat{n} \times \mathbf{E}$ and $\hat{n} \times \mathbf{H}$ are associated with impressed surface magnetic current $\mathbf{M}_s$ and impressed surface electric current $\mathbf{J}_s$, respectively, the above can be thought of having these equivalence impressed surface currents radiating via the dyadic Green's function.

Again, notice that (13.4.14) is derived via the use of (13.4.9), but no boundary condition has yet been imposed on $\overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}')$ on S even though we have given a closed form solution for the free-space case previously. The above is similar to the outside-in equivalence theorem we have derived in Section 13.1.2 using a Gedanken experiment . Now we have a mathematical derivation of the same theorem!

Next, if we require the additional boundary condition that $\hat{n} \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = 0$ for $\mathbf{r} \in S$, this corresponds to a point source located at $\mathbf{r}'$ radiating via the dyadic Green's function, producing a field at $\mathbf{r}$, in the presence of a PEC surface. Then (13.4.14) becomes

$$\mathbf{E}(\mathbf{r}') = -\oint_S dS\, \hat{n} \times \mathbf{E}(\mathbf{r}) \cdot \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}'), \qquad \mathbf{r}' \in V \tag{13.4.15}$$

for, using the scalar triple product rule, it could be shown that $\hat{n} \times \mathbf{H} \cdot \overline{\mathbf{G}} = \mathbf{H} \cdot \hat{n} \times \overline{\mathbf{G}}$ implying that the second term in (13.4.14) is zero on a PEC surface due to the boundary condition we impose on $\hat{n} \times \overline{\mathbf{G}}$. On the other hand, if we have a PMC surface, and we require that $\hat{n} \times \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = 0$ for $\mathbf{r} \in S$, then (13.4.14) becomes

$$\mathbf{E}(\mathbf{r}') = j\omega\mu \oint_S dS\, \hat{n} \times \mathbf{H}(\mathbf{r}) \cdot \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}'), \qquad \mathbf{r}' \in V \tag{13.4.16}$$

Equations (13.4.15) and (13.4.16) state that $\mathbf{E}(\mathbf{r}')$ is determined if either $\hat{n} \times \mathbf{E}(\mathbf{r})$ or $\hat{n} \times \mathbf{H}(\mathbf{r})$ is specified on S. This is in agreement with the uniqueness theorem. These are the mathematical expressions of the PEC and PMC equivalence problems we have considered previously in Sections 13.2 and 13.3.

In (13.4.14), (13.4.15), and (13.4.16), the closed bounding surface S still includes the surface S_{inf}. Therefore, the dyadic Green's functions in (13.4.15) and (13.4.16) are for a closed cavity since boundary conditions are imposed on S for them. We need to prove next that for a homogeneous open region, the integration over S_{inf} vanishes.

The equations so far derived do not require that the dyadic Green's function $\overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}')$ to be translational invariant (or function of only $\mathbf{r} - \mathbf{r}'$), but it has to be a solution of (13.4.9). Thus, the equation for $\overline{\mathbf{G}}$ used in (13.4.15), using (13.4.12), can also be written as

$$\overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = \frac{1}{k^2}[\nabla \times \nabla \times \bar{\mathbf{I}}\, g(\mathbf{r}, \mathbf{r}') - \bar{\mathbf{I}}\, \delta(\mathbf{r} - \mathbf{r}')], \tag{13.4.17}$$

In the above, we can use the BAC-CAB formula to simplify the double curl operator $\nabla \times \nabla \times$

Taking the curl of the above, we have

$$\nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = \nabla \times \bar{\mathbf{I}}\, g(\mathbf{r}, \mathbf{r}'). \tag{13.4.18}$$

In the back-substitution exercise we did earlier, the dyadic Green's function we have used so far is kosher if the scalar Green's function satisfies the following equation

$$(\nabla^2 + k^2) g(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}') \tag{13.4.19}$$

without having to satisfy the radiation condition.[12] Hence, if we can solve the above equation for both bounded and unbounded regions, the dyadic Green's function defined in (13.4.12) is valid both for a homogeneous region that is bounded as well as unbounded. For an unbounded, homogeneous medium, the scalar Green's function, satisfying the radiation condition, has a closed form and is simply

$$g(\mathbf{r},\mathbf{r}') = \frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|}$$

as previously derived. Then, (13.4.14), for $\mathbf{r}' \in V$ and $\mathbf{r}' \neq \mathbf{r}$, becomes, after going through the detail derivation, we can show that

$$\mathbf{E}(\mathbf{r}') = -\oint_S dS\ \hat{n}\times\mathbf{E}(\mathbf{r})\cdot\left(\nabla\times g(\mathbf{r},\mathbf{r}')\bar{\mathbf{I}}\right) + \frac{j}{\omega\epsilon}\oint_S dS\,\hat{n}\times\mathbf{H}(\mathbf{r})\cdot\nabla\times\left(\nabla\times g(\mathbf{r},\mathbf{r}')\bar{\mathbf{I}}\right). \tag{13.4.20}$$

Next, we can assume that $g(\mathbf{r},\mathbf{r}')$ is translationally invariant and can be replaced by $g(\mathbf{r}-\mathbf{r}')$. Then using that $\nabla g(\mathbf{r}-\mathbf{r}') = -\nabla' g(\mathbf{r}-\mathbf{r}')$, the above can be rewritten as

$$\mathbf{E}(\mathbf{r}') = \nabla'\times\oint_S dS\, g(\mathbf{r}-\mathbf{r}')\,\hat{n}\times\mathbf{E}(\mathbf{r}) + \frac{j}{\omega\epsilon}\nabla'\times\nabla'\times\oint_S dS\, g(\mathbf{r}-\mathbf{r}')\,\hat{n}\times\mathbf{H}(\mathbf{r}). \tag{13.4.21}$$

The above can be applied to the geometry in Figure 13.7 where $\mathbf{r}'$ is enclosed in S and S_{inf}. However, the integral over S_{inf} vanishes by virtue of the radiation condition as for (13.4.5).[13] Then, (13.4.21) relates the field outside S at $\mathbf{r}'$ in terms of only the equivalence surface currents $\mathbf{M}_s = \mathbf{E}\times\hat{n}$ and $\mathbf{J}_s = \hat{n}\times\mathbf{H}$ on S. This is similar to the inside-out problem in the equivalence theorem (see Section 13.1.1). It is also related to the fact that if the radiation condition is satisfied, then the field outside of the source region is uniquely satisfied. Therefore, this is also related to the uniqueness theorem.

13.5 Some Math Details

In the above, the equation (13.4.20) follows from (13.4.14). If we post-multiply it by an arbitrary vector $\mathbf{a}$, we can extract a term that resembles

$$\mathbf{A}\cdot(\nabla\times\bar{\mathbf{I}}g)\cdot\mathbf{a} \tag{13.5.1}$$

where $\mathbf{A} = \hat{n}\times\mathbf{E}$. Using the definition that $\bar{\mathbf{I}}g\cdot\mathbf{a} = g\mathbf{a}$, then we have

$$\mathbf{A}\cdot(\nabla\times\bar{\mathbf{I}}g)\cdot\mathbf{a} = -\mathbf{A}\cdot\nabla'\times(g\mathbf{a}) = -(\nabla'\times g\mathbf{a})\cdot\mathbf{A} = -(\nabla'\times g\mathbf{A})\cdot\mathbf{a} \tag{13.5.2}$$

The first term of the above equation (13.4.21) follows from the use of the above identity after cancelling $\mathbf{a}$ in the above equality.

[12] For instance, $g = g_h + g_s$, where g_h is the homogeneous solution and g_s satisfies the above equation with the singularity on the right-hand side.

[13] It is to be noted that the integral over S_{inf} does not vanish because the field is vanishingly small, but the cancellation of the two terms in (13.4.21).

For the second term, we obtain a term from (13.4.14) and (13.4.20) that resembles

$$\mathbf{B} \cdot (\nabla \times \nabla \times g\bar{\mathbf{I}}) \cdot \mathbf{a} \tag{13.5.3}$$

where $\mathbf{B} = \hat{n} \times \mathbf{H}$. It can be shown to be equivalent to

$$\mathbf{B} \cdot (\nabla \times \nabla \times g\bar{\mathbf{I}}) \cdot \mathbf{a} = \mathbf{B} \cdot (\nabla \times \nabla \times (g\mathbf{a}) = \nabla' \times \nabla' \times (g\mathbf{a}) \cdot \mathbf{B} = \nabla' \times \nabla' \times g\mathbf{B} \cdot \mathbf{a} \tag{13.5.4}$$

After cancelling the vector $\mathbf{a}$ from the above equality, we obtain that $\mathbf{B} \cdot (\nabla \times \nabla \times g\bar{\mathbf{I}}) = \nabla' \times \nabla' \times g\mathbf{B}$. Using the above, we can get the second term in (13.4.21).

Exercises for Lecture 13

Problem 13-1:

(i) Using reciprocity theorem, show that an impressed current source on a PEC surface cannot radiate any field.

(ii) A dyad is defined in physics as juxtaposed of two 3 vectors, e.g., **ab** where **a** and **b** are three-component vectors called 3 vectors in physics. In matrix algebra, this is written as $\mathbf{a} \cdot \mathbf{b}^t$ called an outer product (this outer product is denoted $\mathbf{a} \bigotimes \mathbf{b}$ in the math literature as well), where $\mathbf{a}^t \cdot \mathbf{b}$ is called an inner product. In physics, an inner product is just written $\mathbf{a} \cdot \mathbf{b}$. Assuming that **a** and **b** are independent of each other but not orthogonal, even though the dyad **ab** behaves like a 3×3 matrix, it only has one nonzero eigenvector with one nonzero eigenvalue. Find this eigenvector.

It can be shown that a dyad has at most one nonzero eigenvector. The number of nonzero eigenvalues of a matrix is the rank of the matrix. A dyad has a rank of at most 1. (This exercise teaches you what a dyad is.)

(iii) On the other hand, a dyadic is a superposition of dyads. The dyadic Green's function is "homomorphic" to the scalar Green's function, albeit with more complicated vector algebra. By direct back substitution, show that the solution to the following equation

$$\nabla \times \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') - k^2 \, \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = \overline{\mathbf{I}} \, \delta(\mathbf{r} - \mathbf{r}'). \tag{E13.1}$$

is given by

$$\overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = \left(\overline{\mathbf{I}} + \frac{\nabla \nabla}{k^2} \right) g(\mathbf{r}\, vr') \tag{E13.2}$$

In the above, the dyadic $\overline{\mathbf{I}}$ is similar to an identity matrix because $\overline{\mathbf{I}} \cdot \mathbf{a} = \mathbf{a}$. Discuss how you would write $\overline{\mathbf{I}}$ in a three space. You can write $\overline{\mathbf{I}}$ in cartesian, cylindrical, or spherical coordinates.

(iv) Given the vector wave equation for the electric field has a source term, namely that

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r}) - k^2 \, \mathbf{E}(\mathbf{r}) = -j\omega\mu \mathbf{J} \tag{E13.3}$$

By the principle of linear superposition, show that the solution to the above equation can be written as

$$\mathbf{E}(\mathbf{r}) = -j\omega\mu \iiint d\mathbf{r}' \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') \cdot \mathbf{J}(\mathbf{r}') \tag{E13.4}$$

Problem 13-2: This problem is on Huygens' principle for electromagnetic fields in the lecture notes.

(i) Verify the identity (13.4.13) and then derive the expression (13.4.14).

(ii) Explain why (13.4.15) and (13.4.16) are the mathematical expressions of the equivalence problems shown in Sections 13.2 and 13.3.

(iii) Go through the details, derive (13.4.20) for the free space case.

(iv) The surface integral is (13.4.21) is over the surface S and S_{inf} in Figure 13.7. Show that the surface integral over S_{inf} evaluates to zero by cancellation of two terms.

Part II

Transmission Lines, Waves in Layered Media, Waveguides, and Cavity Resonators

Chapter 14

Circuit Theory Revisited

Circuit theory is one of the most successful and often used theories in electrical engineering. Its success is mainly due to its simplicity: it can capture the essence of the physics of highly complex circuits and structures, which is very important in the computer and micro-chip industry (or the IC design industry). In electrical engineering, we distill complicated concepts into the simplest form possible. Simplicity rules! Now, having understood electromagnetic theory in its full glory, it is prudent to revisit circuit theory and study its relationship to electromagnetic theory [57, 33, 68, 34]. Circuit theory can be regarded as the poor-man's version of electromagnetics theory. It is simple, and yet, can be equally powerful in many ways.

The two most important laws in circuit theory are Kirchoff current law (KCL) and Kirchhoff voltage law (KVL) [14, 55]. These two laws predate Maxwell's equations and were driven by the need for telegraphy technology then. But they are derivable from the current continuity equation, which follows from Ampere's law, and from Faraday's law.

14.1 Kirchhoff Current Law (KCL)

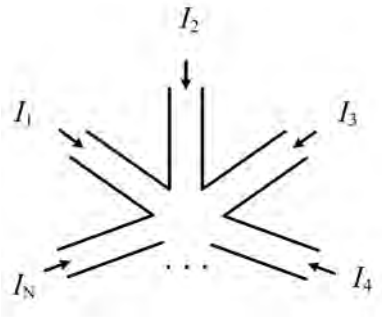


Figure 14.1: Schematic showing the derivation of Kirchhoff current law. All currents flowing into a node must add up to zero. This is reminiscent of the current continuity equation $\nabla \cdot \mathbf{J} = -\partial_t \varrho$, which in the static limit, becomes $\nabla \cdot \mathbf{J} = 0$.

Kirchhoff current law follows from the static Ampere's law that states that

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{14.1.1}$$

By taking the divergence of the above, we arrive at

$$\nabla \cdot \mathbf{J} = 0 \tag{14.1.2}$$

from which KCL can be derived. The above is also the current continuity equation in the static limit.

First, we assume that all currents are flowing into a node as shown in Figure 14.1. By integrating the above current continuity equation over a volume containing the node, it is easy to show that

$$\sum_{i}^{N} I_i = 0 \tag{14.1.3}$$

which is the statement of KCL. This is shown for the schematic of Figure 14.1.

14.2 Kirchhoff Voltage Law (KVL)

Kirchhoff voltage law is the consequence of Faraday's law in the static limit. For the truly static case when $\omega = 0$, it is

$$\nabla \times \mathbf{E} = 0 \tag{14.2.1}$$

The above implies that $\mathbf{E} = -\nabla\Phi$ or that scalar potential theory prevails. From this, we can deduce that

$$-\oint_C \mathbf{E}\cdot\mathbf{dl} = 0 \tag{14.2.2}$$

For statics, the statement that $\mathbf{E} = -\nabla\Phi$ also implies that we can define a voltage drop between two points, a and b to be

$$V_{ba} = -\int_a^b \mathbf{E}\cdot\mathbf{dl} = \int_a^b \nabla\Phi\cdot\mathbf{dl} = \Phi(\mathbf{r}_b) - \Phi(\mathbf{r}_a) = V_b - V_a \tag{14.2.3}$$

The equality $\int_a^b \nabla\Phi\cdot\mathbf{dl} = \Phi(\mathbf{r}_b) - \Phi(\mathbf{r}_a)$ can be understood by expressing this integral in one dimension along a straight line segment, or that

$$\int_a^b \frac{d}{dx}\Phi\cdot\mathbf{dx} = \Phi(\mathbf{r}_b) - \Phi(\mathbf{r}_a) \tag{14.2.4}$$

A curved line can be thought of as a concatenation of many small straight line segments.

The above derivations of KCL and KVL from Maxwell's equations are surrealistically simple largely due to the work of Maxwell [112], and later, the Maxwellians [42] who distilled Maxwell's work further so that it can be easily absorbed by us.

As shall be shown later, to be exact, $\mathbf{E} = -\nabla\Phi - \partial/\partial t\mathbf{A}$, where the second term is due to the induction effect due to $\partial_t\mathbf{B}$ in Faraday's law. Therefore, when we ignore the induction effect, this is only valid in the low frequency or long wavelength limit,[1] or that the dimension over which the above is applied is very small so that retardation effect can be ignored.

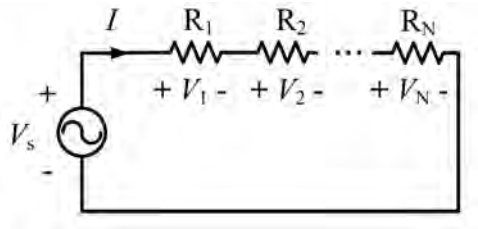


Figure 14.2: Kichhoff voltage law where the sum of all voltages around a loop is zero, which is the consequence of static Faraday's law.

A good way to remember the above formula is that if $V_b > V_a$ then the potential at b is higher than that at a. Since $\mathbf{E} = -\nabla\Phi$, then the electric field points from point a to point b: Electric field always points from the point of higher potential to point of lower potential. Faraday's law when applied to the static case for a closed loop of resistors shown in Figure 14.2 gives Kirchhoff voltage law (KVL), or that

$$\sum_i^N V_j = 0 \tag{14.2.5}$$

[1] These two concepts are synonymous in this course.

Notice that the voltage drop across a resistor is always positive, since the voltages to the left of the resistors in Figure 14.2 are always higher than the voltages to the right of the resistors. This implies that internal to the resistor, there is always an electric field that points from the left to the right. Therefore, the potential on the left side is always higher than that on the right side. A resistor impedes the flow of current, and hence, positive charges accumulate on the left side with negative charges on the right side. An electric field thus points from the left to the right as shown in Figure 14.3.

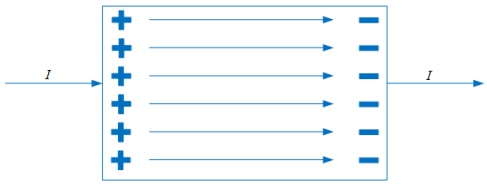


Figure 14.3: The schematic of the field inside a resistor. Due to charge accumulation, the potential on the left side is always higher than that on the right side. An electric field thus points from the left to the right inside the resistor.

If one of the voltage drops is due to a voltage source, it can be modeled by a negative resistor as shown in Figure 14.4.[2] The voltage drop across a negative resistor is opposite to that of a positive resistor. As we have learn from the Poynting's theorem, negative resistor gives out energy instead of dissipates energy. Remember that the complex power dissipated at one point in space is given by $\mathbf{E} \cdot \mathbf{J}^* = \sigma |\mathbf{E}|^2$ after letting $\mathbf{J} = \sigma \mathbf{E}$.

[2]This seemingly simple concept was later awarded the Nobel prize to Esaki [54] to explain how a microwave transistor worked.

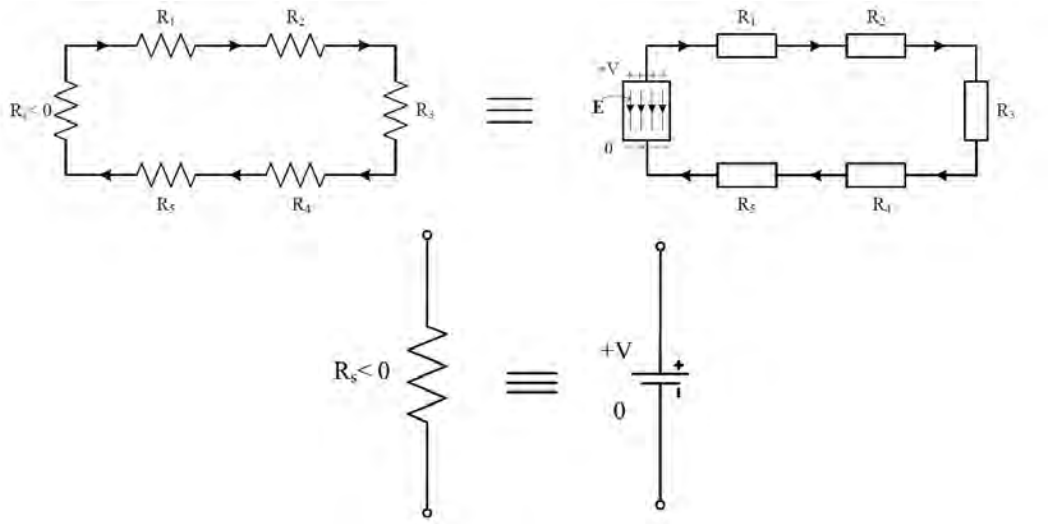


Figure 14.4: A voltage source can also be modeled by a negative resistor.

14.2.1 Faraday's Law and the Flux Linkage Term

Faraday's law for the time-varying **B** flux case is

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{14.2.6}$$

In the static or long-wavelength limit, the right-hand side of the above can be ignored, but it can be amplified with the use of an inductor.

Writing the above in integral form, one gets

$$-\oint_C \mathbf{E} \cdot \mathbf{dl} = \frac{d}{dt} \int_s \mathbf{B} \cdot \mathbf{dS} \tag{14.2.7}$$

For simplicity, we apply the above to a loop shown in Figure 14.5, or a loop C that goes from a to b to c to d to a but the contour C does not go through the wires of the inductor. We can further assume that this loop is very small compared to wavelength so that potential theory that $\mathbf{E} = -\nabla\Phi$ applies.[3] Furthermore, we assume that the frequency is low such that this loop C has little or no magnetic flux through it so that the right-hand side of the above can be approximated by zero, or Faraday's law becomes

$$-\oint_C \mathbf{E} \cdot \mathbf{dl} = 0 \tag{14.2.8}$$

[3]You will later learn that the exact expression for the electric field is $\mathbf{E} = -\nabla\Phi - \partial_t \mathbf{A}$, where $\mathbf{A}$ is the magnetic vector potential. $\mathbf{A}$ is related to the magnetic field but since the frequency is low, or that there are no magnetic field in the region considered, electric potential Φ suffices to describe the electric field $\mathbf{E}$.

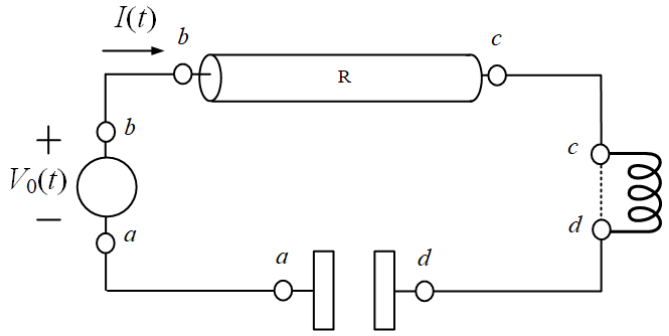


Figure 14.5: The Kirchhoff voltage law (KVL) for a circuit loop consisting of resistor, inductor, and capacitor can also be derived from Faraday's law at low frequency. The KVL can be derived by integrating Faraday's law around the loop defined by *abcda*. This loop does not have any magnetic flux linkage. But the detail of V_{cd} and how it relates to the flux linkage in the inductor is discussed in the text. At low frequency, the flux linkage term in Faraday's law, and the displacement current term in the generalized Ampere's law are less important, unless their effects can be amplified by using inductors and capacitors. Thus the full glory of Maxwell's equation is not lost in circuit theory (redrawn from original Figure 4.2 in Ramo et al) [33].

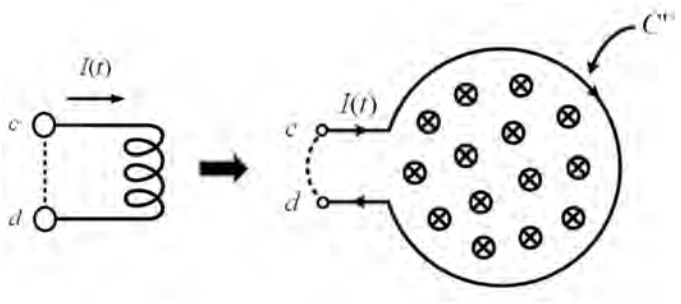


Figure 14.6: The voltage-current relation of an inductor can be obtained by unwrapping an inductor coil, and then calculating its flux linkage.

Since this loop does not go through the inductor, but goes directly from c to d, there is no flux linkage in this loop. Or $\mathbf{E} = -\nabla\Phi$, and thus

$$-\int_a^b \mathbf{E}\cdot\mathbf{dl} - \int_b^c \mathbf{E}\cdot\mathbf{dl} - \int_c^d \mathbf{E}\cdot\mathbf{dl} - \int_d^a \mathbf{E}\cdot\mathbf{dl} = 0 \tag{14.2.9}$$

Inside the source or the battery, it is assumed that the electric field points opposite to the direction of integration **dl**, and hence the first term on the left-hand side of the above is positive and equal

to $V_0(t)$, while the other terms are negative. Writing out the above more explicitly, after using (14.2.3), we have

$$V_0(t) + V_{cb} + V_{dc} + V_{ad} = 0 \tag{14.2.10}$$

Notice that in the above, in accordance to (14.2.3), $V_b > V_c$, $V_c > V_d$, and $V_d > V_a$. Therefore, V_{cb}, V_{dc}, and V_{ad} are all negative quantities but $V_0(t) > 0$. We will study the contributions to each of the terms, the inductor, the capacitor, and the resistor more explicitly next.

14.2.2 Inductor—Flux Linkage Amplifier

An inductor is a lumped element that is used to amplify the flux linkage term in Faraday's law. To find the voltage current relation of an inductor, we apply Faraday's law to a closed loop C' formed by dc and the inductor coil shown in the Figure 14.6 where, for simplicity, we have unwrapped the solenoid into a larger loop. Assume that the inductor is made of a perfect conductor, so that the electric field $\mathbf{E}$ in the wire is zero. Then the only contribution to the left-hand side of Faraday's law is the integration from point d to point c, the only place in the loop C' where $\mathbf{E}$ is not zero. We assume that outside the loop in the region between c and d, potential theory applies; namely, we can let $\mathbf{E} = -\nabla\Phi$. Now, we can connect V_{dc} in the previous equation to the flux linkage to the inductor. When the voltage source attempts to drive an electric current into the loop, Lenz's law (1834)[4] comes into effect, essentially generating an opposing voltage. The opposing voltage gives rise to charge accumulation at d and c, and therefore, a low frequency electric field at the gap at dc.

To this end, we form a new C' that goes from d to c, and then continue onto the wire that leads into the inductor. But this new loop will contain the flux $\mathbf{B}$ generated by the inductor current. Thus

$$\oint_{C'} \mathbf{E}\cdot\mathbf{dl} = \int_d^c \mathbf{E}\cdot\mathbf{dl} = -V_{dc} = -\frac{d}{dt}\int_{S'} \mathbf{B}\cdot\mathbf{dS} \tag{14.2.11}$$

As mentioned before, since the wire is a PEC, the integration around the loop C' is only nonzero from d to c. In the above, $\int_{S'} \mathbf{B}\cdot\mathbf{dS}$ is the flux linkage, which should be linearly proportional to I, the current. The inductance L is then defined as the flux linkage per unit current, or

$$L = \left[\int_{S'} \mathbf{B}\cdot\mathbf{dS}\right]/I \tag{14.2.12}$$

So the voltage in (14.2.11) is then

$$V_{dc} = \frac{d}{dt}(LI) = L\frac{dI}{dt} \tag{14.2.13}$$

since L is time independent.

Had there been a finite resistance in the wire of the inductor, then the electric field is non-zero inside the wire. Taking this into account, we have

$$\oint_{C'} \mathbf{E}\cdot\mathbf{dl} = R_L I - V_{dc} = -\frac{d}{dt}\int_S \mathbf{B}\cdot\mathbf{dS} \tag{14.2.14}$$

[4]Lenz's law can also be explained from Faraday's law (1831).

Consequently,

$$V_{dc} = R_L I + L\frac{dI}{dt} \tag{14.2.15}$$

Thus, to account for the loss of the coil, we can add a resistor in the equation. The above becomes simpler in the frequency domain, namely

$$V_{dc} = R_L I + j\omega L I \tag{14.2.16}$$

14.2.3 Capacitance—Displacement Current Amplifier

The capacitance is the proportionality constant between the charge Q stored in the capacitor, and the voltage V applied across the capacitor, or $Q = CV$. Then

$$C = \frac{Q}{V} \tag{14.2.17}$$

From the current continuity equation, one can easily show that in Figure 14.7,

$$I = \frac{dQ}{dt} \tag{14.2.18}$$

by charge conservation. Furthermore, using that $Q = CV_{da}$, we have

$$I = \frac{d}{dt}(CV_{da}) = C\frac{dV_{da}}{dt}$$

where C is time independent. Integrating the above equation, one gets

$$V_{da}(t) = \frac{1}{C}\int_{-\infty}^{t} I dt' \tag{14.2.19}$$

The above looks quite cumbersome in the time domain, but in the frequency domain, it becomes

$$I = j\omega C V_{da} \tag{14.2.20}$$

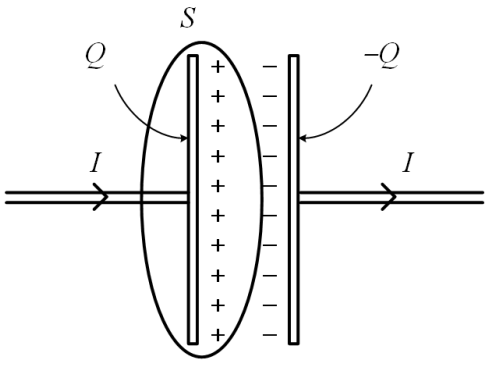


Figure 14.7: Schematic showing the calculation of the capacitance of a capacitor.

14.3 Resistor

The electric field is not zero inside the resistor as electric field is needed to push electrons through it. As discussed in Section 8.3, a resistor is a medium where collision of the electrons with the lattice dominates. As is well known, here Ohm's law prevails, and

$$\mathbf{J} = \sigma \mathbf{E} \tag{14.3.1}$$

where σ is the conductivity of the medium. From this, we deduce that $V_{cb} = V_c - V_b$ is a negative number given by

$$V_{cb} = -\int_b^c \mathbf{E} \cdot \mathbf{dl} = -\int_b^c \frac{\mathbf{J}}{\sigma} \cdot \mathbf{dl} \tag{14.3.2}$$

where we assume a uniform current $\mathbf{J} = \hat{l} I/A$ in the resistor where $\hat{l}$ is a unit vector pointing in the direction of current flow in the resistor. We can assumed that I is a constant along the length of the resistor, and thus, with $\mathbf{dl} = \hat{l} dl$, $\mathbf{J} \cdot \mathbf{dl} = I dl/A$, implying that

$$V_{cb} = -\int_b^c \frac{I dl}{\sigma A} = -I \int_b^c \frac{dl}{\sigma A} = -IR \tag{14.3.3}$$

where[5]

$$R = \int_b^c \frac{dl}{\sigma A} = \int_b^c \frac{\rho dl}{A} \tag{14.3.4}$$

Again, for simplicity, we assume long wavelength or low frequency in the above derivation. The above formula is valid approximately even when the cross-sectional area A and resistivity ρ are functions of l.

Inductors, capacitors, and resistors are known as lumped elements because they are usually much smaller than wavelength, but interesting physics occurs inside them. By so doing, circuit theory can be made very simple with far ranging impact.

14.4 Generalized KCL and KVL, the Power of Phasors

Phasor technique allows us to generalize KCL and KVL to include the displacement current and induction voltage easily. To do this for generalized KCL, we rewrite Ampere's law in the frequency domain as

$$\nabla \times \mathbf{H} = j\omega \mathbf{D} + \mathbf{J} = \mathbf{J}_D + \mathbf{J} = \mathbf{J}_T \tag{14.4.1}$$

where the displacement current $\mathbf{J}_D = j\omega \mathbf{D}$. Taking the divergence of the above equation, we have

$$\nabla \cdot \mathbf{J}_T = 0 \tag{14.4.2}$$

where $\mathbf{J}_T = \mathbf{J}_D + \mathbf{J}$ includes the displacement current. KCL can be written easily for $\mathbf{J}_T$, which is very simple in the frequency domain. The above is the root to the generalized KCL. Note that here, the current $\mathbf{J}_T$ is complex.

[5]The resisitivity $\rho = 1/\sigma$ where ρ has the unit of ohm-m, while σ has the unit of siemen/m.

Similarly, we rewrite Faraday's law in the frequency domain as

$$\nabla \times \mathbf{E} = -j\omega\mathbf{B} = -j\omega\nabla \times \mathbf{A} \tag{14.4.3}$$

where we have defined $\mathbf{B} = \nabla \times \mathbf{A}$ as before. From the above, we move all terms to the left-hand side, then we can define a new electric field $\underset{\sim}{\mathbf{E}} = \mathbf{E} + j\omega\mathbf{A}$ such that

$$\nabla \times \underset{\sim}{\mathbf{E}} = 0 \tag{14.4.4}$$

The above is the root to the generalized KVL. Now we can now define $\underset{\sim}{\mathbf{E}} = \mathbf{E} + j\omega\mathbf{A} = -\nabla\Phi$. We can integrate the above around a closed loop to get

$$\oint_C \mathbf{E} \cdot d\mathbf{l} + j\omega \oint_C \mathbf{A} \cdot d\mathbf{l} = 0 \tag{14.4.5}$$

Assuming that the closed loop in the first term is made of the PEC wire C' and air region from d to c. Thus the integration around the cloded loop is nonzero only in the air region from d to c. Therefore,

$$V_{dc} = \int_d^c \mathbf{E} \cdot d\mathbf{l} = -j\omega \oint_C \mathbf{A} \cdot d\mathbf{l} = -j\omega \int_S (\nabla \times \mathbf{A}) \cdot d\mathbf{S} = -j\omega \int_S \mathbf{B} \cdot d\mathbf{S} = -j\omega L \tag{14.4.6}$$

The left-hand side is the voltage between points c and d, while the right-hand side is the flux linkage term responsible for generating this counter induction voltage. If we add a resistor in the loop, then the above can be converted to

$$V_{dc} = -R - j\omega L \tag{14.4.7}$$

indicating the V_{dc} is a complex voltage.

14.5 Some Remarks

We have looked at the definition of inductor L and capacitor C in the above. But clever engineering is driven by heuristics: it is better, at times, to look at inductors and capacitors as energy storage devices, rather than flux linkage and charge storage devices.

Another important remark is that even though circuit theory is simpler that Maxwell's equations in its full glory, not all the physics is amiss in it. The physics of the induction term in Faraday's law and the displacement current term in generalized Ampere's law are still retained and can be amplified by inductor and capacitor, respectively. In fact, wave physics is still retained in circuit theory: one can make slow wave structure out a series of inductors and capacitors. The lumped-element model of a transmission line is an example of a slow-wave structure design that will be studied in the next chapter. Since the wave is slow, it has a smaller wavelength, and as shall be seen, resonators can be made smaller: We see this in the LC tank circuit which is a much smaller resonator in wavelength with $L/\lambda \ll 1$ compared to a microwave cavity resonator for instance. Therefore, circuit design is great for miniaturization. The short coming is that inductors and capacitors are made from material media, and generally have higher losses than air or vacuum.

14.6 Energy Storage Method for Inductor and Capacitor

As aforementioned, often time, it is more expedient to think of inductors and capacitors as energy storage devices. This enables us to identify stray (also called parasitic) inductances and capacitances more easily. This manner of thinking allows for an alternative way of calculating and understanding inductances and capacitances as well [33].

The energy stored in an inductor is due to its energy storage in the magnetic field, and it is alternatively written, according to circuit theory, as

$$W_m = \frac{1}{2}LI^2 \rightarrow L = \frac{2W_m}{I^2} \tag{14.6.1}$$

Therefore, it is simpler to think that an inductance exists whenever there is stray magnetic field to store magnetic energy. A piece of wire carries a current that produces a magnetic field enabling energy storage in the magnetic field. Hence, with this insight, we can understand why a piece of wire in fact behaves like a small inductor, which is non-negligible at high frequencies: Stray inductances occur whenever there are stray magnetic fields.

By the same token, a capacitor can be thought of as an electric energy storage device rather than a charge storage device. The energy stored in a capacitor, from circuit theory, is

$$W_e = \frac{1}{2}CV^2 \rightarrow C = \frac{2W_e}{V^2} \tag{14.6.2}$$

Therefore, whenever stray electric field exists, one can think of stray capacitances as we have seen in the case of fringing field capacitances in a microstrip line.

14.7 Finding Closed-Form Formulas for Inductance and Capacitance

Closed form formulas are important as they give us engineering and physical insight. Finding closed form solutions for inductors and capacitors is a difficult endeavor. As in solving Maxwell's equations or the waveguide problems, only certain simple geometries are amenable to closed form solutions. Even a simple circular loop does not have a closed form solution for its inductance L. If we assume a uniform current on a circular loop, in theory, the magnetic field can be calculated using Bio-Savart law that we have learnt before, namely that

$$\mathbf{H}(\mathbf{r}) = \int \frac{I(\mathbf{r}')\mathbf{dl}' \times \hat{R}}{4\pi R^2} \tag{14.7.1}$$

But the above cannot be evaluated in closed form except in terms of complicate elliptic integrals [113, 114]. Thus it is simpler to just measure the inductance.

However, if we have a solenoid as shown in Figure 14.8, an approximate formula for the inductance L can be found if the fringing field at the end of the solenoid can be ignored. The inductance can be found using the flux linkage method [33, 31]. Figure 14.9 shows the schematic used to find the approximate inductance of this inductor.

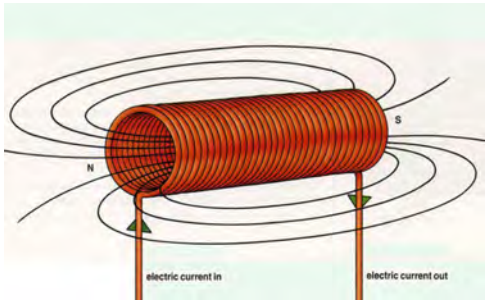


Figure 14.8: The flux-linkage method is used to estimate the inductance of a solenoid (courtesy of SolenoidSupplier.Com).

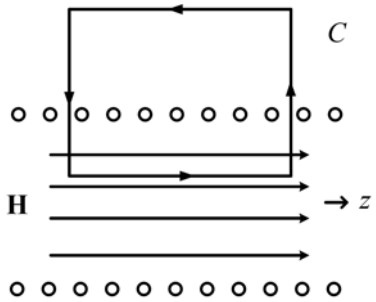


Figure 14.9: Finding the inductor flux linkage approximately by assuming the magnetic field is uniform inside a long solenoid. This approximation is increasingly better as the solenoid becomes longer.

The capacitance of a parallel plate capacitor can be found by solving a boundary value problem (BVP) for electrostatics as shown in Section 3.3.5. The electrostatic BVP for capacitor involves Poisson's equation and Laplace equation which are scalar equations [49]. Finding the correct formula for the capacitor as shown in Figure 14.10 involving fringing field effect can be an exhaustive exercise [115]. Alternatively, variational expressions can be used to find the lower and upper bounds of capacitors using, for example, Thomson's theorem [116] together with numerical methods. [6]

[6]There are many variational formulas for capacitance some of which are discussed in [116][p. 53]. It seems that nature (or God) always tries to minimize something in seeking the solution. He is a minimalist.

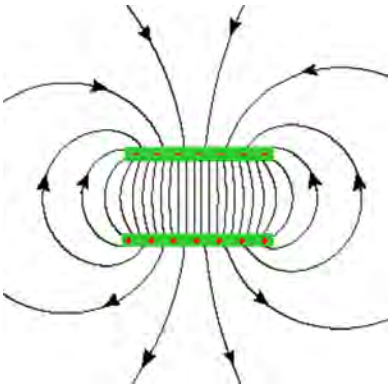


Figure 14.10: Nominally, the field in between two parallel plates in a capacitor is non-uniform. When the parallel plate is large, the ball-park value of the capacitor can be estimated by assuming a uniform field in between them. The correction to this simple formula incorporating fringing fields requires some tour-de-force analysis [115] (courtesy of quora.com).

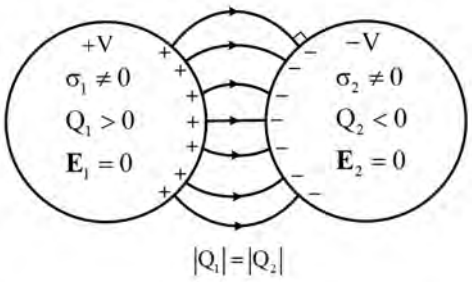


Figure 14.11: The capacitance between two charged conductors can be found by solving a boundary value problem (BVP) involving Laplace equation as discussed in 3.3.5. If the conductors are of odd shapes, then numerical methods are needed.

Assume a geometry of two conductors charged to $+V$ and $-V$ volts as shown in Figure 14.11. Surface charges will accumulate on the surfaces of the conductors. Using Poisson's equations, and Green's function for Poisson's equation, one can express the potential in between the two conductors as due to the surface charges density $\sigma(\mathbf{r})$. It can be expressed as (using (3.3.22) of Chapter 3)

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon}\int_S dS' \frac{\sigma(\mathbf{r}')}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{14.7.2}$$

where $S = S_1 + S_2$ is the union of two surfaces S_1 and S_2. Since Φ has values of $+V$ and $-V$ on

the two conductors, we require that

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon}\int_S dS' \frac{\sigma(\mathbf{r}')}{4\pi|\mathbf{r}-\mathbf{r}'|} = \begin{cases} +V, & \mathbf{r}\in S_1 \\ -V, & \mathbf{r}\in S_2 \end{cases} \tag{14.7.3}$$

In the above, $\sigma(\mathbf{r}')$, the surface charge density, is the unknown yet to be sought and it is embedded in an integral. But the right-hand side of the equation is known. Hence, this equation is also known as an integral equation where the unknown to be sought is embedded inside the integral. The integral equation can be solved by numerical methods that shall be discussed later.

Having found $\sigma(\mathbf{r})$, then it can be integrated to find Q, the total charge on one of the conductors. Since the voltage difference between the two conductors is known, the capacitance can be found as $C = Q/(2V)$. Here, $2V$ is assumed because it is the voltage difference between the two objects.

14.8 Importance of Circuit Theory in IC Design

The clock rate of computer circuits has peaked at about 3 GHz due to the resistive loss, or the I^2R loss. At this frequency, the wavelength is about 10 cm. Since transistors and circuit components are shrinking due to the compounding effect of Moore's law, most components, which are of nanometer dimensions, are much smaller than the wavelength. Thus, most of the physics of electromagnetic signal in a microchip circuit can be captured by using circuit theory.

Figure 14.12 shows the schematic and the cross section of a computer chip at different levels: with the transistor at the bottom-most level. The signals are taken out of a transistor by XY lines at the middle level that are linked to the ball-grid array at the top-most level of the chip. And then, the signal leaves the chip via a package. At the chip level, these structures are nanometer-size and are much smaller than the wavelength,[7] they are usually modeled by lumped R, L, and C elements when retardation effect can be ignored. If retardation effect is needed, it is usually modeled by a transmission line. The retardation effect is important at the package level where the dimensions of the components are sizeable compared to wavelength.

A process of parameter extraction where computer software or field solvers (software that solves Maxwell's equations numerically, e.g., HFSS [117, 118] and PEEC [103]) are used to extract these lumped-element parameters. Finally, a computer chip is modeled as a network involving a large number of transistors, diodes, and R, L, and C elements. Subsequently, a very useful and powerful commercial software called SPICE (Simulation Program with Integrated-Circuit Emphasis) [119], is a computer-aided software, that helps to solve for the voltages and currents in this network.

[7] Circuit theory applies when the dimension of a structure is much smaller than a wavelength.

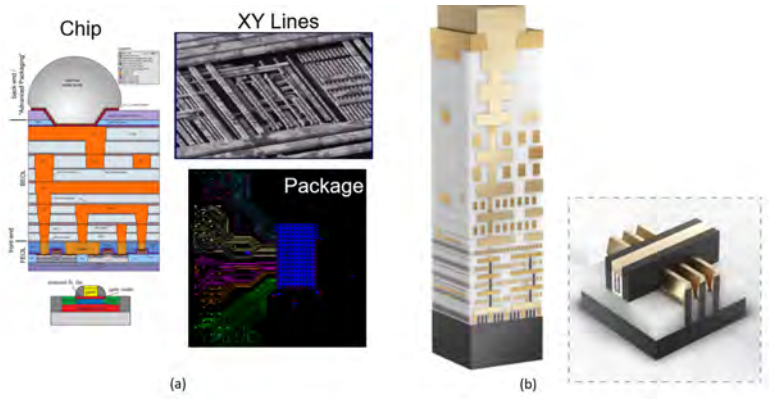


Figure 14.12: (a) Cross section of a chip (top left) and the XY lines in the chip (top right), and the interconnects in the package needed to take the signal out of the chip (bottom right). (b) A cross-sectional view of the modern design of a chip with FinFET transistors at the bottom, which switch faster and are smaller. Nowadays, transistors on chip number in billions. Clever engineering is needed to extract the humongous amount of data from these transistors (courtesy of Wikipedia and K. Radhakrishnan, Intel).

Initially, SPICE software was written primarily to solve circuit problems. But the SPICE software now has many capabilities, including modeling of transmission lines (that shall be discussed next) for microwave engineering, which are important for modeling retardation effects. Figure 14.13 shows a graphical user interface (GUI) of an RF-SPICE that allows the modeling of transmission line with a Smith chart interface.

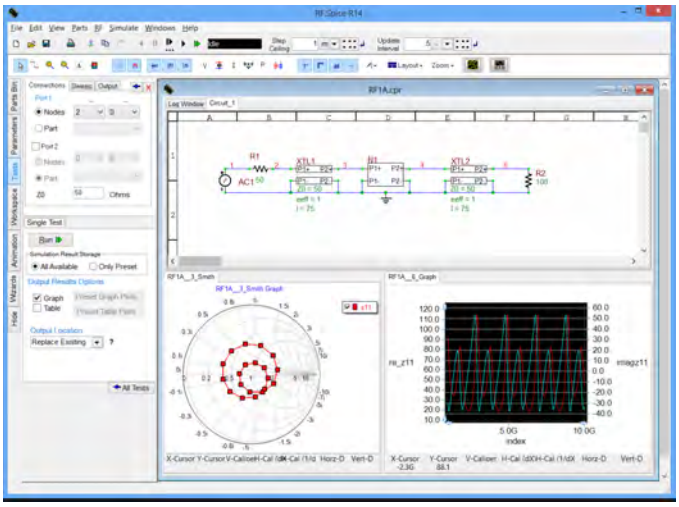


Figure 14.13: SPICE is also used to solve RF problems. A transmission line, to be discussed in the next lecture, is used in combination with circuit theory to account for retardation effects in a computer circuit (courtesy of EMAG Technologies Inc.).

14.9 Decoupling Capacitors and Spiral Inductors

Decoupling capacitor is an important part of modern computer chip design. They can regulate voltage supply on the power delivery network (PDN) of the chip as they can remove high-frequency noise and voltage fluctuation from a circuit as shown in Figure 14.14. Figure 14.15 shows a 3D IC computer chip where decoupling capacitors are integrated into its design.

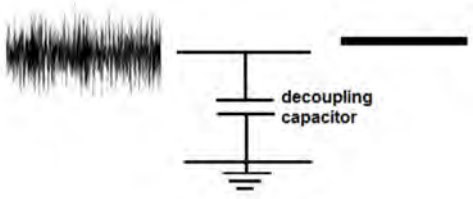


Figure 14.14: A decoupling capacitor is essentially a low-pass filter allowing low-frequency signal to pass through, while high-frequency signal is short-circuited (courtesy learningaboutelectronics.com).

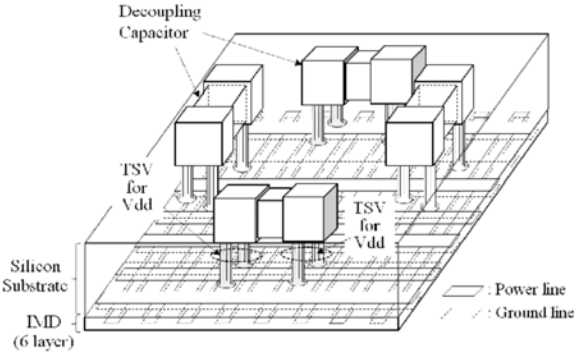


Figure 14.15: Modern computer chip design is 3D and is like an "urban jungle" like Manhattan or Hong Kong. There are different levels in the chip and they are connected by through silicon vias (TSV). IMD stands for inter-metal dielectrics. One can see different XY lines serving as power and ground lines (courtesy of Semantic Scholars). Because of the complexity of modern chip designs, and the common electromagnetic interference between the components, signal integrity (SI) and power integrity (PI) are important issues in their designs.

Inductors are also indispensable in IC design, as they can be used as a high frequency choke. However, designing compact inductor on a chip is still a challenge. Spiral inductors are used because of their planar structure and ease of fabrication. However, miniaturizing inductor is a difficult frontier research topic [120].

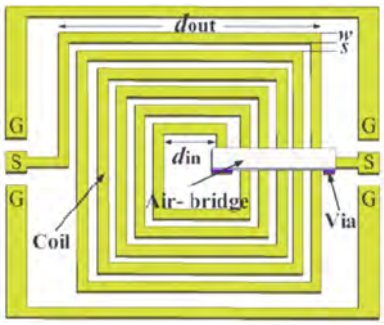


Figure 14.16: Spiral inductors are difficult to build on a chip, but by using laminal structure, it can be integrated into the IC fabrication process (courtesy of Quan Yuan, Research Gate). The fabrication of circuit components has to be commensurate with the photolithography and chemical processes used.

14.10 Why the 3 GHz Barrier?

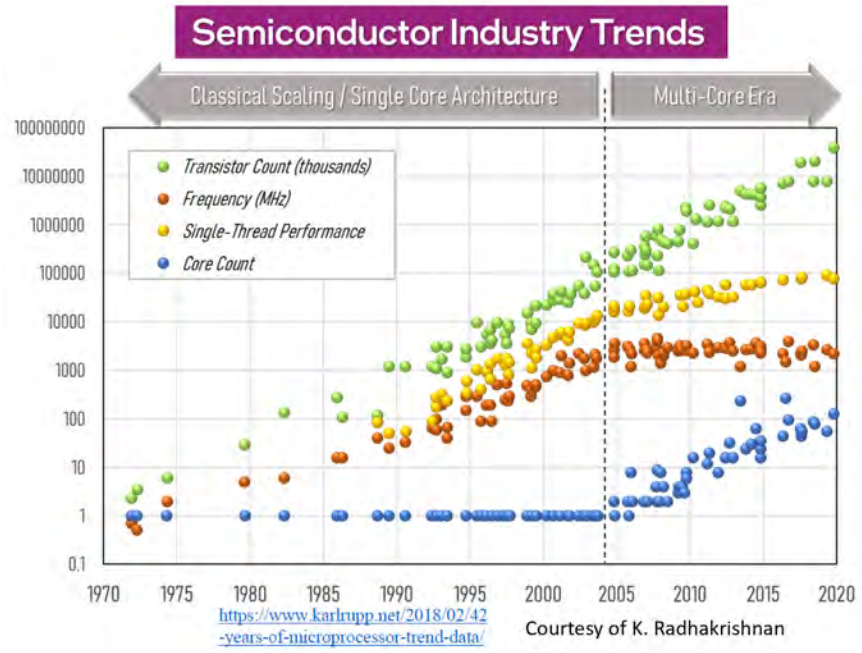


Figure 14.17: The semiconductor trends reveal the self-fulfilling prophesy of Moore's law. Ingenious engineering and 3D architecture of the microchip design allow the transistor and core counts to improve the performance of the microchip. Also, the clock rate (frequency) has not increased in the last two decades due to joule heating to be explained next.

Moore's law projects that the packing density of transistors will double every 1.5 years. This trend has continued unabated until the modern time. It has allowed each of us to carry the equivalence of the supercomputers of yesteryears in our pockets. Some of these cell phones have over 100 billion transistors in them. The most advanced have trillions of transistors in them. This unrelenting growth is truly the gift of God to us! It has given us humongous amount of cloud storage, and free email accounts. This growth is sometimes called the self-fulfilling prophesy of Moore's law. Because the semiconductor industry wants to see this growth, engineers become more ingenious in their chip designs. As you can see, chips are now three-dimensional via the use of through-silicon vias (TSV).

Despite the exciting growth of the microchip industry, the clock rate of computer chips has not increase very much: it has stagnated around 3 GHz. Thus, to improve the performance of microchips, core counts have increased to counter the stagnation in the clock rate (see Figure 14.17). There is a reason why the clock rate of CPU has not improved even though transistors now can switch at 800 GHz. The reason is joule heating as explained simply by Figure 14.18.

The gate voltage of a MOSFET is used to switch it on and off. It can be modeled simply by a gate capacitor. The transistors of a microchip are connected to the outside world by interconnects. The power delivery network (PDN) can be modeled by a voltage source connected to the gate by a wire as shown in Figure 14.18. The resistive loss on the wire can be modeled by a resistor R as shown. As the clock rate goes up, the frequency goes up, and the gate capacitance appears to be more like a short circuit, increasing the driving current I. Hence, the I^2R loss increases, giving rise to joule heating.[8] This joule heating, and the need to dissipate and manage this heat is the reason why the clock rate has not risen in our computers in the last two decades.

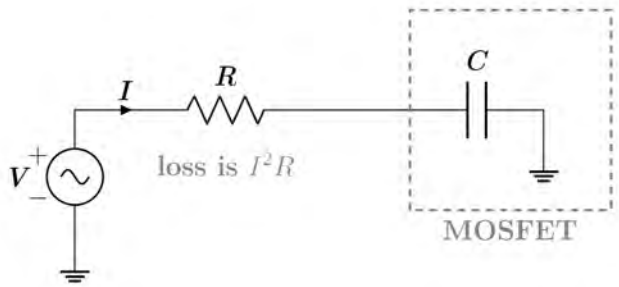


Figure 14.18: MOSFET driven by a power delivery network via the interconnects can be modeled simply as a gate capacitance that models the MOSFET transistor. The metal loss in the interconnects can be modeled as a resistor R in series with a voltage source V. As the frequency increases, more current flows through the capacitance, and hence, through the resistor, increasing the I^2R loss giving rise to joule heating.

14.11 When is Circuit Theory Valid?

Before we end this lecture, it will be good to ponder for a moment to ask, "When is circuit theory valid?" We will use dimensional analysis to help us answer this question. Faraday's law and Ampere's law are given as

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{14.11.1}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J} \tag{14.11.2}$$

$$\tag{14.11.3}$$

Potential theory for which $\mathbf{E} = -\nabla\Phi$ follows from Faraday's law if we can drop the $\partial_t\mathbf{B}$ term in Faraday's law. Then KVL can be derived. Also, if we can drop the $\partial_t\mathbf{D}$ term in Ampere's law, then we have $\nabla \times \mathbf{H} = \mathbf{J}$, which implies that $\nabla \cdot \mathbf{J} = 0$ from which KCL can be derived. Then if we can drop the time derivative terms in the above, we can apply KVL and KCL.

[8] I am indebted to Paul Y.S Cheung of HKU for sharing this insight with me.

For a geometry where the dimensions of the devices are of $O(L)$, the fields have to vary on the length scale of L in order to satisfy the boundary conditions around the devices. Thus we can argue that $\frac{\partial}{\partial x} \sim 1/L$ and similarly in all the other directions such as the y and z directions.

Thus if we can argue that

$$\frac{\partial}{c\partial t} \ll \frac{\partial}{\partial x} \tag{14.11.4}$$

where c is the velocity of light.[9] For a sinusoidal signal, $\partial_t \sim O(\omega)$. The above is equivalent to

$$\frac{\omega}{c} \ll \frac{1}{L} \tag{14.11.5}$$

Since ω/c is the wavenumber which is $\frac{2\pi}{\lambda}$, the above is the same as

$$\frac{2\pi}{\lambda} \ll \frac{1}{L}, \text{ or } 2\pi L \ll \lambda \tag{14.11.6}$$

Therefore, if the dimensions L of the devices we are dealing with are much smaller that the wavelength λ, we can apply potential theory or circuit theory.

[9] We cannot compare apples with oranges, so we have to make the units on both sides of the inequality the same.

Exercises for Lecture 14

Problem 14-1: Circuit theory is often regarded as a subset of electromagnetic theory, but it is very simple, and hence popular. By amplifying the displacement current term using capacitors and flux linkage term using inductors, the full glory of electromagnetic theory is embedded in circuit theory.

(i) Derive Kirchhoff current law and Kirchhoff voltage law from electromagnetic theory.

(ii) Use the energy storage method: Find the inductance of the solenoid and show that it is the same as that obtained by the flux linkage method. Hint: Ramo et al has a discussion of this.

(iii) Use the energy storage method: Find the capacitance of two parallel plate capacitor, and show that it is the same as solving Laplace's equation as a boundary value problem.

Problem 14-2:

(i) Using your undergraduate electrical engineering knowledge, show that the magnetic field energy storage in an inductor is given by

$$W_m = \frac{1}{2} L I^2$$

as in (14.6.1).

(ii) Similarly, show that the electric field energy storage in a capacitor is given by

$$W_e = \frac{1}{2} C V^2$$

as in (14.6.2).

Chapter 15

Transmission Lines

Transmission lines represent one of the most important electromagnetic technologies. The reason being that they can be described by simple theory, similar to circuit theory. As such, the theory is within the grasp of most practicing electrical engineers. Moreover, transmission line theory fills the gap in the physics of circuit theory: Circuit theory alone cannot describe wave phenomena, but when circuit theory is augmented with transmission line theory, wave phenomena with its corresponding wave physics such as retardation and time delay start to emerge.

Even though circuit theory has played an indispensable role in the development of the computer chip industry, eventually, circuit theory has to be embellished by transmission line theory, so that high-speed circuits can be designed. Retardation effect, which causes time delay, clock skew, and phase shift, can be modeled simply using transmission lines. Nowadays, commercial circuit solver software such as SPICE[1] have the capability of including transmission line as an element in its modeling.

[1]This is an acronym for a package "simulation program with integrated circuit emphasis" that came out of U. Cal., Berkeley [119].

15.1 Transmission Line Theory

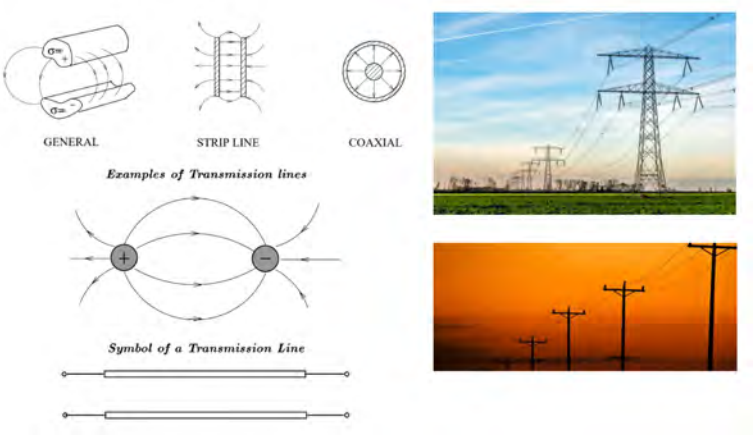


Figure 15.1: Various kinds of transmission lines. Schematically, all of them can be modeled by two parallel wires. On the right are pictures of a power transmission line, and a telephone line. Due to technology advancement, telephone lines are a rare sight now (courtesy of Lister-Communications and Istockphoto.com).

Transmission lines were the first electromagnetic waveguides ever invented. The were driven by the needs in telegraphy technology. It is best to introduce transmission line theory from the viewpoint of circuit theory, which is elegant and one of the simplest theories of electrical engineering. This theory is also discussed in many textbooks and lecture notes. Transmission lines are so important in modern day electromagnetic engineering, that most engineering electromagnetics textbooks would be incomplete without introducing the topics related to them [53, 121, 57, 85, 33, 68, 34, 36, 90, 47].

Circuit theory is robust and is not sensitive to the detail shapes of the components involved such as capacitors or inductors. Moreover, many transmission line problems cannot be analyzed simply when the full form of Maxwell's equations is used,[2] but approximate solutions can be obtained using circuit theory. We have seen previously that circuit theory is an approximation of electromagnetic field theory when the wavelength is very long (or the frequency is low): the longer the wavelength, the better is the approximation [53]. Hence, in long-wavelength limit, transmission line theory can be approximated by circuit theory.

Examples of transmission lines are shown in Figure 15.1. The symbol for a transmission line is just two pieces of parallel wires, but in practice, these wires need not be parallel as shown in Figure 15.2.

[2] Usually called full-wave analysis.

Figure 15.2: A twisted pair transmission line where the two wires are not parallel to each other (courtesy of slides by A. Wadhwa, A.L. Dal, N. Malhotra [122]). The winding of the wires in series opposition cancels the linkage of the transmission line to low frequency external magnetic flux.

Circuit theory also explains why waveguides such as transmission lines can be made sloppily when wavelength is long or the frequency low. For instance, in the long-wavelength limit, we can make twisted-pair waveguides with abandon, and yet they still work well (see Figure 15.2). Hence, it is simplest and best to first explain the propagation of electromagnetic signal on a transmission line using circuit analysis.

15.1.1 Time-Domain Analysis

We will start with performing the time-domain analysis of a simple, infinitely long transmission line. Remember that two pieces of metal can accumulate attractive positive and negative charges between them, giving rise to electric fields that start with positive charges and end with negative charges. The stored energy in the electric field gives rise to capacitive effect in the line which can be modeled by capacitances. Moreover, a piece of wire carrying a current generates a magnetic field, and hence, yields stored energy in the magnetic field. The stored magnetic field energy gives rise to inductive effect in the line which can be modeled by inductances. These stored energies are the sources of the capacitive and inductive effects.

But these capacitive and inductive effects are distributed over the spatial dimension of the transmission line. Therefore, it is helpful to think of the two pieces of metal as consisting of small segments of metal concatenated together. Each of these segments will have a small inductance, as well as a small capacitive coupling between them. Therefore, we can model two pieces of metal with a distributed lumped element model as shown in Figure 15.3. For simplicity, we assume the other conductor to be a ground plane, so that it need not be approximated with lumped elements.

In the transmission line, the voltage $V(z,t)$ and the current $I(z,t)$ are functions of both space z and time t, but we will model the space variation of the voltage and current with discrete step approximations. The voltage varies from node to node while the current varies from branch to branch of the lumped-element model.

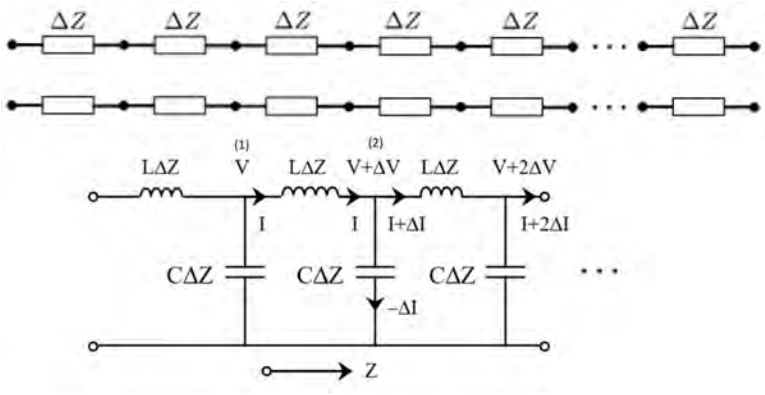


Figure 15.3: A long transmission line can be replaced by a concatenation of many short transmission lines. For each pair of short wires, there are capacitive coupling between them. Furtheremore, when current flows in the wire, magnetic field is generated making them behave like an inductor. Therefore, the transmission line can be replaced by a lumped-element approximation as shown. The lumped elements have inductances given by $L\Delta z$ and capacitances given by $C\Delta z$, distributed over the line. This is also known as the distributive model of the transmission line. There were many lumped element models of a transmission line, but the final one that captures the physics correctly is due to Heaviside [42].

Telegrapher's Equations

First, we recall that the V-I relation of an inductor is

$$V_0 = L_0 \frac{dI_0}{dt} \tag{15.1.1}$$

where L_0 is the inductor, V_0 is the time-varying voltage drop across the inductor, and I_0 is the current through the inductor. Then using this relation between nodes 1 and 2 in Figure 15.3, we have

$$V - (V + \Delta V) = L\Delta z \frac{\partial I}{\partial t} \tag{15.1.2}$$

The left-hand side is the voltage drop across then series inductor, while the right-hand side follows from the aforementioned V-I relation of an inductor, but we have replaced $L_0 = L\Delta z$. Here, L is the inductance per unit length (line inductance) of the transmission line. And $L\Delta z$ is the incremental inductance due to the small segment of metal of length Δz. In the above, we assume that $V = V(z,t)$ and $I = I(z,t)$, so that time derivative is replaced by partial time derivative. Then the above (15.1.2) can be simplified to

$$\Delta V = -L\Delta z \frac{\partial I(z,t)}{\partial t} \tag{15.1.3}$$

where ΔV is the incremental voltage drop between the two nodes 1 and 2.

Next, we make use of the V-I relation for a single capacitor, which is

$$I_0 = C_0 \frac{dV_0}{dt} \tag{15.1.4}$$

where C_0 is the capacitor, I_0 is the current through the capacitor, and V_0 is a time-varying voltage drop across the capacitor. Thus, applying this relation at node 2 in Figure 15.3 gives the incremental shunt current to be

$$-\Delta I = C\Delta z \frac{\partial}{\partial t}(V + \Delta V) \approx C\Delta z \frac{\partial V}{\partial t} \tag{15.1.5}$$

where C is the capacitance per unit length, and $C\Delta z$ is the incremental capacitance between the small piece of metal and the ground plane. In the above, we have used Kirchhoff current law to surmise that the current through the shunt capacitor is $-\Delta I$, where $\Delta I = I(z + \Delta z, t) - I(z, t)$. In the last approximation in (15.1.5), we have dropped a term involving the product of Δz and ΔV, since it will be very small or second order in magnitude.

In the limit when $\Delta z \to 0$, one gets from (15.1.3) and (15.1.5) that

$$\frac{\partial V(z,t)}{\partial z} = -L \frac{\partial I(z,t)}{\partial t} \tag{15.1.6}$$

$$\frac{\partial I(z,t)}{\partial z} = -C \frac{\partial V(z,t)}{\partial t} \tag{15.1.7}$$

The above are the *telegrapher's equations*.[3] They are two coupled first-order equations, and can be converted into second-order equations easily by eliminating one of the two unknowns. Therefore,

$$\frac{\partial^2 V}{\partial z^2} - LC \frac{\partial^2 V}{\partial t^2} = 0 \tag{15.1.8}$$

$$\frac{\partial^2 I}{\partial z^2} - LC \frac{\partial^2 I}{\partial t^2} = 0 \tag{15.1.9}$$

The above are wave equations that we have previously studied, where the velocity of the wave is given by

$$v = \frac{1}{\sqrt{LC}} \tag{15.1.10}$$

Furthermore, if we assume that

$$V(z,t) = V_0 f_+(z - vt), \qquad I(z,t) = I_0 f_+(z - vt) \tag{15.1.11}$$

corresponding to a right-traveling wave, they can be verified to satisfy (15.1.6) and (15.1.7) as well as (15.1.8) and (15.1.9) by back substitution.

[3] They can be thought of as the distillation of the Faraday's law, $\nabla \times \mathbf{E} = -\partial_t \mathbf{B}$, and Ampere's law, $\nabla \times \mathbf{H} = \partial_t \mathbf{D}$, from Maxwell's equations without the source term. Their simplicity gives them an important role in engineering electromagnetics. One can think of them as poor man's Maxwell's equations.

Consequently, we can easily show that for the right-traveling wave,

$$\frac{V(z,t)}{I(z,t)} = \frac{V_0}{I_0} = \sqrt{\frac{L}{C}} = Z_0 \tag{15.1.12}$$

where Z_0 is a constant independent of z, which is the characteristic impedance of the transmission line. The above ratio is only true for one-way traveling wave, in this case, one that propagates in the $+z$ direction.

For a wave that travels in the negative z direction, in a similar manner, we can let,

$$V(z,t) = V_0 f_-(z+vt), \qquad I(z,t) = I_0 f_-(z+vt) \tag{15.1.13}$$

one can easily show by the same token that

$$\frac{V(z,t)}{I(z,t)} = \frac{V_0}{I_0} = -\sqrt{\frac{L}{C}} = -Z_0 \tag{15.1.14}$$

Again, Z_0 is a constant independent of z.

Time-domain analysis is very useful for transient analysis of transmission lines, especially when nonlinear elements are coupled to the transmission line.[4] Another major strength of transmission line model is that it is a simple way to introduce time-delay (also called retardation) in a simple circuit model.[5] As we saw when we studied the solution to the wave equation: solutions at different times are just the time-delayed version of the original solution.

Time Delay and Velocity of Light

Time delay is a wave propagation effect, and it is harder to incorporate into circuit theory or a pure circuit model consisting of R, L, and C only. In circuit theory, where the wavelength is assumed very long, Laplace's equation is usually solved, which is equivalent to Helmholtz equation with infinite wave velocity, namely,

$$\lim_{c\to\infty} \nabla^2\Phi(\mathbf{r}) + \frac{\omega^2}{c^2}\Phi(\mathbf{r}) = 0 \quad \Longrightarrow \quad \nabla^2\Phi(\mathbf{r}) = 0 \tag{15.1.15}$$

From the above, we see that Helmholtz equation becomes Laplace's equation when the velocity of light c is infinite. Hence, events in Laplace's equation happen instantaneously. In other words, in circuit theory, where Laplace's equation is usually involved, we assume that the velocity of the wave is infinite, and there is no retardation effect. This is only true or a good approximation when the size of the structure is small compared to wavelength.

15.1.2 Frequency-Domain Analysis—the Power of Phasor Technique Again!

As we have seen in previous lectures, the frequency-domain analysis greatly simplifies the analysis of many complicated phenomena. This was especially true in our analysis of conductive media,

[4] Remember that we can only use frequency domain technique or Fourier transform for linear time-invariant systems.

[5] By a simple circuit model, we mean a model that has lumped elements such as R, L, and C as well as a transmission line element.

and frequency dispersive media as in the Drude-Lorentz-Sommerfeld model etc. As such, frequency domain analysis is very popular as it makes the transmission line equations very simple–one just replace $\partial/\partial t \to j\omega$. Moreover, generalization to a lossy system is quite straightforward. Furthermore, for linear time invariant systems, the time-domain signals can be obtained from the frequency-domain data by performing a Fourier inverse transform since phasors and Fourier transforms of a time variable are just related to each other by a constant (see Section 6.2).

The telegrapher's equations (15.1.6) and (15.1.7) then in frequency domain become

$$\frac{d}{dz}V(z,\omega) = -j\omega LI(z,\omega) \tag{15.1.16}$$

$$\frac{d}{dz}I(z,\omega) = -j\omega CV(z,\omega) \tag{15.1.17}$$

The above gives the notion that the change in the voltage $V(z,\omega)$ on a transmission line is proportional to the line impedance $j\omega L$ times the current $I(z,\omega)$. Similar notion can be said of the second equation above.

The corresponding 1D Helmholtz equations can be derived from the above, and they are then

$$\frac{d^2V}{dz^2} + \omega^2 LCV = 0 \tag{15.1.18}$$

$$\frac{d^2I}{dz^2} + \omega^2 LCI = 0 \tag{15.1.19}$$

The above are second order ordinary differential equations (ODE), and the general solutions to the above are

$$V(z) = V_+ e^{-j\beta z} + V_- e^{j\beta z} \tag{15.1.20}$$

$$I(z) = I_+ e^{-j\beta z} + I_- e^{j\beta z} \tag{15.1.21}$$

where the wavenumber $\beta = \omega\sqrt{LC}$. This is similar to what we have seen previously for plane waves in the one-dimensional wave equation in free space, where

$$E_x(z) = E_{0+} e^{-jk_0 z} + E_{0-} e^{jk_0 z} \tag{15.1.22}$$

Here, $k_0 = \omega\sqrt{\mu_0\epsilon_0}$. We see much similarity between (15.1.20), (15.1.21), and (15.1.22).

To see the solution in the time domain, we let the phasor $V_\pm = |V_\pm|e^{j\phi_\pm}$ in (15.1.20), and the voltage signal above can then be converted back to the time domain using the key formula in phasor technique as

$$V(z,t) = \Re e\{V(z,\omega)e^{j\omega t}\} \tag{15.1.23}$$

$$= |V_+|\cos(\omega t - \beta z + \phi_+) + |V_-|\cos(\omega t + \beta z + \phi_-) \tag{15.1.24}$$

As can be seen, the first term corresponds to a right-traveling wave, while the second term is a left-traveling wave.

Furthermore, if we assume only a one-way traveling wave to the right by letting $V_- = I_- = 0$, then it can be shown that, for a right-traveling wave, using (15.1.16) or (15.1.17)

$$\frac{V(z)}{I(z)} = \frac{V_+}{I_+} = \sqrt{\frac{L}{C}} = Z_0 \tag{15.1.25}$$

where Z_0 is the characteristic impedance. The above is the same as that in the time domain we have derived previously.[6] Since Z_0 is real, it implies that the phasors[7] $V(z)$ and $I(z)$ are in phase.

Similarly, applying the same process for a left-traveling wave only, by letting $V_+ = I_+ = 0$, then

$$\frac{V(z)}{I(z)} = \frac{V_-}{I_-} = -\sqrt{\frac{L}{C}} = -Z_0 \tag{15.1.26}$$

In other words, for the left-traveling waves, the voltage and current are 180° out of phase. This is similar to the time domain case we have before.

15.2 Lossy Transmission Line

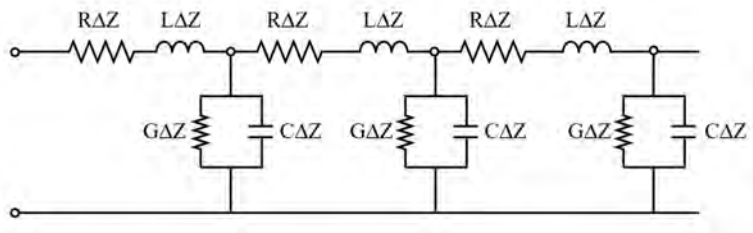


Figure 15.4: In a lossy transmission line, series resistance can be added to the series inductance, and the shunt conductance can be added to the shunt susceptance of the capacitor. This problem is "homomorphic" to the lossless case in the frequency domain.

The phasor technique is empowered by, first, that the algebra for complex numbers is the same as that of real numbers. Second, resistors and conductances are replaced by impedances and admittances in the frequency domain. By exploiting "homomorphism", this makes the solution involving a network of impedances and admittances analogous to the network of resistances and conductances. The power of frequency domain analysis is revealed in the study of lossy transmission lines. The previous analysis, which is valid for lossless transmission line, can be easily generalized to the lossy case in the frequency domain. In using frequency domain and phasor technique, impedances will become complex numbers as shall be shown.

To include loss, we use the lumped-element model as shown in Figure 15.4. One thing to note is that $j\omega L$ is actually the series line impedance of the lossless transmission line, while $j\omega C$ is the shunt line admittance of the same line. First, we can rewrite the expressions for the telegrapher's equations in (15.1.16) and (15.1.17) in terms of series line impedance and shunt line admittance to arrive at

$$\frac{d}{dz}V = -ZI \tag{15.2.1}$$

[6] One can think that this ratio is independent of frequency, and hence, is valid both in the time-domain as well as the frequency domain.

[7] We will neglect to denote phasors by under-tilde, as they are implied by the context.

$$\frac{d}{dz}I = -YV \tag{15.2.2}$$

where $Z = j\omega L$ and $Y = j\omega C$. The above can be easily generalized to the lossy case as shall be shown.

The geometry in Figure 15.4 is topologically similar to, or "homomorphic"[8] to the lossless case in Figure 15.3. Hence, when lossy elements are added in the geometry, we can surmise that the corresponding telegrapher's equations are similar to those above. But to include loss, we need only to generalize the series line impedance and shunt admittance from the lossless case to lossy case as follows:

$$Z = j\omega L \rightarrow Z = j\omega L + R \tag{15.2.3}$$

$$Y = j\omega C \rightarrow Y = j\omega C + G \tag{15.2.4}$$

where R is the series line resistance, and G is the shunt line conductance. Here, R can be used to model copper loss on a metallic transmission line and G, the dielectric leakage loss in the shunt capacitance. Thus, now Z and Y are the series impedance and shunt admittance, (which are complex numbers rather than being pure imaginary numbers), respectively. We will further exploit the fact that the algebra of complex numbers is the same as the algebra of real numbers. We will refer to this as mathematical "homomorphism". Then, the corresponding Helmholtz equations are

$$\frac{d^2V}{dz^2} - ZYV = 0 \tag{15.2.5}$$

$$\frac{d^2I}{dz^2} - ZYI = 0 \tag{15.2.6}$$

or

$$\frac{d^2V}{dz^2} - \gamma^2 V = 0 \tag{15.2.7}$$

$$\frac{d^2I}{dz^2} - \gamma^2 I = 0 \tag{15.2.8}$$

where $\gamma^2 = ZY$, or that one can also think of $\gamma^2 = -\beta^2$ by comparing with (15.1.18) and (15.1.19). Then the above is "homomorphic" to the lossless case except that now, β is a complex number, indicating that the field is decaying and oscillating as it propagates. As before, the above are second order 1D Helmholtz equations where the general solutions are

$$V(z) = V_+ e^{-\gamma z} + V_- e^{\gamma z} \tag{15.2.9}$$

$$I(z) = I_+ e^{-\gamma z} + I_- e^{\gamma z} \tag{15.2.10}$$

and

$$\gamma = \sqrt{ZY} = \sqrt{(j\omega L + R)(j\omega C + G)} = j\beta \tag{15.2.11}$$

[8] A math term for "similar in math structure". The term is even used in computer science describing a emerging field of homomorphic computing.

Here, $\beta = \beta' - j\beta''$ is now a complex number. In other words,

$$e^{-\gamma z} = e^{-j\beta' z - \beta'' z}$$

is an oscillatory and decaying wave. Or focusing on the voltage case, we have

$$V(z) = V_+ e^{-\beta'' z - j\beta' z} + V_- e^{\beta'' z + j\beta' z} \tag{15.2.12}$$

Again, letting $V_\pm = |V_\pm| e^{j\phi_\pm}$, the above can be converted back to the time domain as

$$V(z,t) = \Re e\{V(z,\omega) e^{j\omega t}\} \tag{15.2.13}$$

$$= |V_+| e^{-\beta'' z} \cos(\omega t - \beta' z + \phi_+) + |V_-| e^{\beta'' z} \cos(\omega t + \beta' z + \phi_-) \tag{15.2.14}$$

The first term corresponds to a decaying wave moving to the right while the second term is also a decaying wave but moving to the left. When there is no loss, or $R = G = 0$, and from (15.2.11), we deduce that $j\beta = j\omega\sqrt{LC}$ is pure imaginary, or that $\beta = \beta'$ and that $\beta'' = 0$.

Notice that for the lossy case, the characteristic impedance, which is the ratio of the voltage to the current for a one-way wave, can similarly be derived using "homomorphism":

$$Z_0 = \frac{V_+}{I_+} = -\frac{V_-}{I_-} = \sqrt{\frac{L}{C}} = \sqrt{\frac{j\omega L}{j\omega C}} \rightarrow Z_0 = \sqrt{\frac{Z}{Y}} = \sqrt{\frac{j\omega L + R}{j\omega C + G}} \tag{15.2.15}$$

The above Z_0 is manifestly a complex number. Here, Z_0 is the ratio of the phasors of the one-way traveling waves, and apparently, their current phasor and the voltage phasor will not be in phase for lossy transmission line.

In the absence of loss, the above again becomes

$$Z_0 = \sqrt{\frac{L}{C}} \tag{15.2.16}$$

the characteristic impedance for the lossless case previously derived.

Exercises for Lecture 15

Problem 15-1:

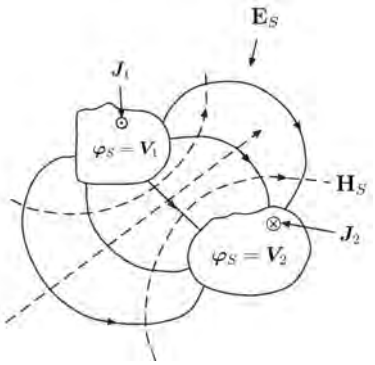


Figure 15.5: The TEM $\mathbf{E}_s$ and $\mathbf{H}_s$ fields of a transmission line.

The TEM field of the transmission line is as shown in Figure 15.5. The electromagnetic fields are purely transverse with no z component. The telegrapher's equations for transmission lines have been derived using circuit theory in the text. Hence, a transmission line can be twisted and bent with abandon, and yet it still works. But the telegrapher's equations can also be derived using field theory, which is more elegant to some people, as shall be shown below:

(i) The fields of a guided TEM (transverse electromagnetic) mode in a transmission line can be written as

$$\mathbf{E} = \mathbf{E}_s e^{-j\beta z}, \qquad \mathbf{H} = \mathbf{H}_s e^{-j\beta z} \tag{15.2.1}$$

where $\mathbf{E}_s$ and $\mathbf{H}_s$ are purely transverse fields with no z components. Show that in order for the above to satisfy Maxwell's equations, then

$$\nabla_s \times \mathbf{E}_s = 0, \qquad \nabla_s \times \mathbf{H}_s = 0 \tag{15.2.2}$$

The subscript s implies transverse to z, and it is also true for the del operator.

(ii) With your knowledge of potential theory, how would you solve the first equation to obtain $\mathbf{E}_s$. You can assume that one of the conductors has voltage V while the other conductor is at ground with zero volt. For simplicity, you may repeat this exercise for the coax cable as has been done in the earlier part of this course.

(iii) Show from Maxwell's equations that once $\mathbf{E}_s$ is known, $\mathbf{H}_s$ can be obtained via the following equation:

$$\frac{\partial}{\partial z}\hat{z} \times \mathbf{E}_s = -j\omega\mu\mathbf{H}_s, \qquad \frac{\partial}{\partial z}\hat{z} \times \mathbf{H}_s = j\omega\mu\mathbf{E}_s \tag{15.2.3}$$

(iv) Can you derive the telegrapher's equation from the above equations? (Hint: Cross product one of the above with $\hat{z}$ to simplify it and then use line integral involving $\mathbf{E}_s$ and $\mathbf{H}_s$ to define the voltage and the current on the conductors (see also [90][p. 40-41].)

Problem 15-2: For the geometry shown in Figure 15.6, it actually does not support at TEM mode, a fact that you shall learn later in the course (also, it actually does not have a closed form solution but we can use a circuit, or quasi-static approximation in finding the transmission line parameters.). Namely, in the long-wavelength limit, we can still define a line capacitance and a line inductance. And then use circuit theory concept to find the phase velocity of a low frequency wave on such a line. The line admittance and impedance of a transmission line can be found by solving an electrostatic problem and a magnetostatic problem, respectively.

(i) The coaxial cable geometry below has inner radius a and outer radius b. It has a lossy dielectric medium between the inner and outer conductor. Find the admittance per unit length that you can substitute into the telegrapher's equations. Since the problem does not have a closed form solution, an approximate solution can be obtained by assuming uniform radial electric field inside the coax. (Also, you may want to think about the lossless problem first where $\sigma_1 = 0$, solve the electrostatics problem, and then use math "homomorphism" between statics and dynamics and phasor technique to obtain the quasi-static dynamic solution.)

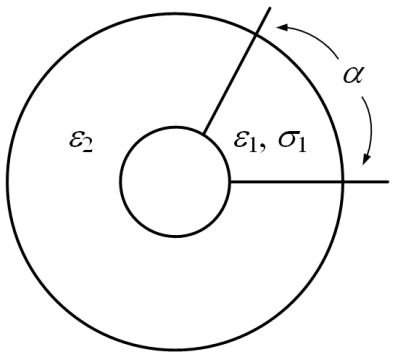


Figure 15.6: A coaxial cable with inhomogeneous, conductive medium inside.

(ii) Since in the static limit, the magnetostatic problem is decoupled from the electrostatic problem, we can solve the magnetostatics problem first. Assume a current I that flows in the inner conductor (or $-I$ in the outer conductor), and that the conductors are lossless. Using the magnetic field you have found before, find the magnetic energy stored per unit length. Knowing that the energy storage for an inductor is $\frac{1}{2}LI^2$, find the inductance per unit length.

(iii) Use your results to find the characteristic impedance of this lossy transmission line, and also the propagation constant γ.

Chapter 16

More on Transmission Lines

As mentioned before, transmission line theory is indispensable in microwave engineering these days. The theory is the necessary augmentation of circuit theory for higher frequency analysis, and it is also indispensable to integrated circuit designers as computer clock rate becomes faster. Over the years, engineers have developed some very useful tools and measurement techniques to expand the design space of circuit designers. We will learn some of these tools in this lecture.[1]

As seen in the previous lecture, the telegrapher's equations are similar to the one-dimensional form of Maxwell's equations, and can be thought of as Maxwell's equations in their simplest form. Therefore, they entail a subset of the physics seen in the full Maxwell's equations. Transmission line is a poor-man's way of incorporating wave physics into engineering designs, without invoking the full bounty of Maxwell's equations.

16.1 Terminated Transmission Lines

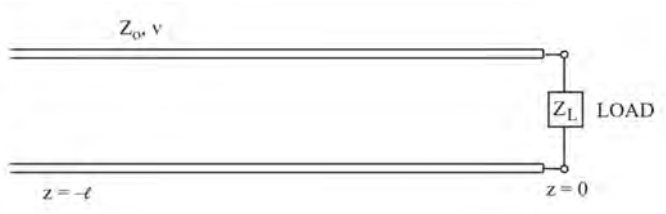


Figure 16.1: A schematic for a transmission line terminated with an impedance load Z_L at $z = 0$.

[1]Some of you may have studied this topic in your undergraduate electromagnetics course. However, this topic is important, and you will have to muster your energy to master this knowledge again:)

For an infinitely long transmission line, the solution consists of the linear superposition of a wave traveling to the right plus a wave traveling to the left. If transmission line is terminated by a load as shown in Figure 16.1, a right-traveling wave will be reflected by the load to generate a left-traveling wave, and in general, the wave on the transmission line will be a linear superposition of the left and right traveling waves.

To simplify the analysis, we will assume that the line is lossless. The generalization to the lossy case is quite straightforward as shall be shown later. Thus, we assume that

$$V(z) = a_+ e^{-j\beta z} + a_- e^{j\beta z} = V_+(z) + V_-(z) \tag{16.1.1}$$

where $V_+(z)$ and $V_-(z)$ are right and left traveling waves, respectively. Here $\beta = \omega/v = \omega\sqrt{LC}$. In the above, in general, $a_+ \neq a_-$. Besides, we assume that the system is linear; or we can define the amplitude of the left-going reflected wave a_- to be linearly related to the amplitude of the right-going or incident wave a_+. In other words, at $z = 0$, we can let

$$V_-(z = 0) = \Gamma_L V_+(z = 0) \tag{16.1.2}$$

where Γ_L is independent of $V_+(z = 0)$, making this relation a linear one. Thus, using the definition of $V_+(z)$ and $V_-(z)$ as implied in (16.1.1), we have

$$a_- = \Gamma_L a_+ \tag{16.1.3}$$

where Γ_L is the termed the reflection coefficient. Hence, (16.1.1) becomes

$$V(z) = a_+ e^{-j\beta z} + \Gamma_L a_+ e^{j\beta z} = a_+ \left(e^{-j\beta z} + \Gamma_L e^{j\beta z}\right) \tag{16.1.4}$$

The corresponding current $I(z)$ on the transmission line is derived using the telegrapher's equations as previously defined. By recalling that

$$\frac{\mathrm{d}V}{\mathrm{d}z} = -j\omega L I$$

then for the general case,

$$I(z) = \frac{a_+}{Z_0}\left(e^{-j\beta z} - \Gamma_L e^{j\beta z}\right) \tag{16.1.5}$$

Notice the sign change in the second term of the above expression.

From (16.1.1) and (16.1.3), one sees that the left-going wave and the right-going wave are linearly proportional to each other. In other words, if we double the right-going wave, the left-going wave doubles. Thus, similar to Γ_L, a general or local reflection coefficient $\Gamma(z)$ at z can be defined (which is a function of z) relating the left-traveling and right-traveling wave at location z such that

$$\Gamma(z) = \frac{V_-(z) = a_- e^{j\beta z}}{V_+(z) = a_+ e^{-j\beta z}} = \frac{a_- e^{j\beta z}}{a_+ e^{-j\beta z}} = \Gamma_L e^{2j\beta z} \tag{16.1.6}$$

where (16.1.3) has been used to relate a_- and a_+ to Γ_L. Of course, $\Gamma(z = 0) = \Gamma_L$. Furthermore, due to the V-I relation at an impedance load, we must have

$$\frac{V(z = 0)}{I(z = 0)} = Z_L \tag{16.1.7}$$

or that using (16.1.4) and (16.1.5) with $z = 0$, the left-hand side of the above can be rewritten, and we have

$$\frac{1+\Gamma_L}{1-\Gamma_L} Z_0 = Z_L, \quad \text{or} \quad \frac{1+\Gamma_L}{1-\Gamma_L} = \frac{Z_L}{Z_0} = Z_{nL} \tag{16.1.8}$$

where Z_{nL} is the normalized load. From the above, we can solve for Γ_L in terms of Z_{nL} to get

$$\Gamma_L = \frac{Z_{nL}-1}{Z_{nL}+1} = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{16.1.9}$$

Thus, given the termination load Z_L and the characteristic impednace Z_0, the reflection coefficient Γ_L can be found, or vice versa. Or given Γ_L, the normalized load impedance, $Z_{nL} = Z_L/Z_0$, can be found.

It is seen that $\Gamma_L = 0$ if $Z_L = Z_0$. Thus a right-traveling wave will not be reflected and the left-traveling is absent. This is the case of a *matched load*. When there is no reflection, all energy of the right-traveling wave will be totally absorbed by the load.

In general, we can define a generalized impedance at $z \neq 0$ to be

$$\begin{aligned} Z(z) = \frac{V(z)}{I(z)} &= \frac{a_+(e^{-j\beta z} + \Gamma_L e^{j\beta z})}{\frac{1}{Z_0} a_+(e^{-j\beta z} - \Gamma_L e^{j\beta z})} \\ &= Z_0 \frac{1+\Gamma_L e^{2j\beta z}}{1-\Gamma_L e^{2j\beta z}} = Z_0 \frac{1+\Gamma(z)}{1-\Gamma(z)} \end{aligned} \tag{16.1.10}$$

where $\Gamma(z)$, the general reflection coefficient defined in (16.1.6) is used in the above. Since the voltage $V(z)$ and $I(z)$ are proportional to the electric field and the magnetic field at location z, this generalized impedance is the ratio of the electric field to the magnetic field at the location z. The above can also be normalized and written as

$$Z_n(z) = Z(z)/Z_0 = \frac{1+\Gamma(z)}{1-\Gamma(z)} \tag{16.1.11}$$

where $Z_n(z)$ is the normalized generalized impedance. Conversely, one can write the above as

$$\Gamma(z) = \frac{Z_n(z)-1}{Z_n(z)+1} = \frac{Z(z)-Z_0}{Z(z)+Z_0} \tag{16.1.12}$$

From (16.1.10) above, one gets

$$Z(z) = Z_0 \frac{1+\Gamma_L e^{2j\beta z}}{1-\Gamma_L e^{2j\beta z}} \tag{16.1.13}$$

One can show that by setting $z = -l$, using (16.1.9) for Γ_L, and after some algebra,

$$Z(-l) = Z_0 \frac{Z_L + jZ_0 \tan\beta l}{Z_0 + jZ_L \tan\beta l} \tag{16.1.14}$$

16.1.1 Short-Circuited Terminations

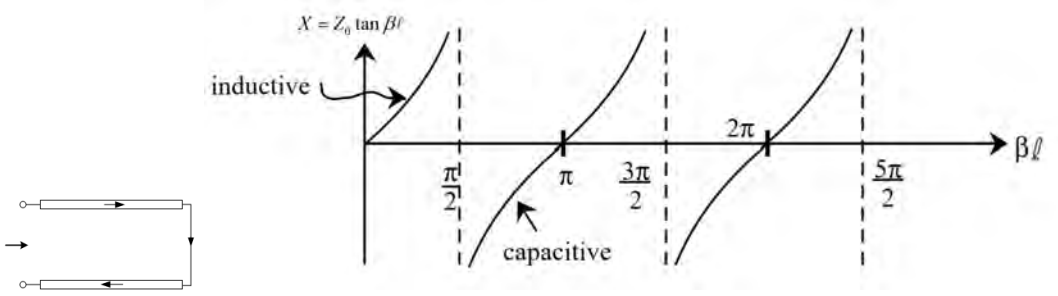


Figure 16.2: The input reactance (X) of a short-circuited transmission line as a function of its length l. The input impedance of the transmission line alternates between being inductive and capacitive as the length varies. A standing wave develops on this short-circuit terminated transmission line. This can make the input impedance look inductive or capacitive depending on if the input port is current or voltage dominated, respectively. Large current flow on a short-circuited line generating stored magnetic energy and hence, the impedance becomes inductive depending on the value of l.

From (16.1.14) above, when we have a short such that $Z_L = 0$, then

$$Z(-l) = jZ_0 \tan(\beta l) = jX \tag{16.1.15}$$

On the short-circuited transmission line, a standing wave develops where the voltage and the current are out of phase with respect to each other. When the current is stronger than the voltage, the magnetic energy stored is larger than the electric energy stored as in the case of an inductor, the input impedance is inductive. However, if one varies the length such that at the input port, the voltage is stronger than the current, the input impedance becomes capacitive. When $\beta l \ll 1$, then $\tan(\beta l) \approx \beta l$, and (16.1.15) becomes

$$Z(-l) \cong jZ_0 \beta l \tag{16.1.16}$$

After using that $Z_0 = \sqrt{L/C}$ and that $\beta = \omega\sqrt{LC}$, (16.1.16) becomes

$$Z(-l) \cong j\omega Ll = j\omega L_{\text{eff}} \tag{16.1.17}$$

The above implies that a short length of transmission line connected to a short as a load looks like an inductor with $L_{\text{eff}} = Ll$, since much current will pass through this short producing a strong magnetic field with stored magnetic energy. Remember here that L is the line inductance, or inductance per unit length.

16.1.2 Open-Circuited Terminations

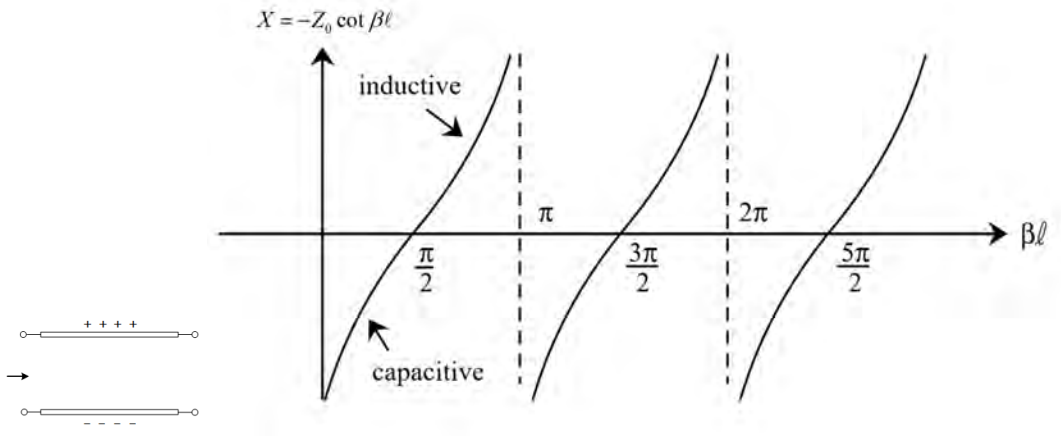


Figure 16.3: The input reactance (X) of an open-circuited transmission line as a function of its length l. Again, a standing wave develops on this open-circuit terminated transmission line. This can make the input impedance look capacitive or inductive depending on the length l.

When we have an open circuit such that $Z_L = \infty$, then from (16.1.14) above

$$Z(-l) = -jZ_0 \cot(\beta l) = jX \tag{16.1.18}$$

Then, when $\beta l \ll 1$, $\cot(\beta l) = 1/\tan(\beta l) \approx 1/\beta l$

$$Z(-l) \approx -j\frac{Z_0}{\beta l} \tag{16.1.19}$$

And then, again using $\beta = \omega\sqrt{LC}$, $Z_0 = \sqrt{L/C}$

$$Z(-l) \approx \frac{1}{j\omega C l} = \frac{1}{j\omega C_{\text{eff}}} \tag{16.1.20}$$

Hence, an open-circuited terminated short length of transmission line appears like an effective capacitor with $C_{\text{eff}} = Cl$. Again, remember here that C is the line capacitance or capacitance per unit length of the transmission line.

As shown in Figure 16.3, the impedance at $z = -l$ is purely reactive when there is no loss, and goes through positive and negative values due to the standing wave set up on the transmission line. Therefore, by changing the length of l, one can make a short-circuited or an open-circuited line look like an inductor or a capacitor depending on its length l. This effect is shown in Figures 16.2 and 16.3. Moreover, the reactance X becomes infinite or zero with the proper choice of the length l. These are resonances or anti-resonances of the transmission line, very much like the resonance of an LC tank circuit. An LC circuit can look like an open or a short circuit at resonances and depending on if they are connected in parallel or in series.

16.2 Smith Chart

In general, from (16.1.13) and (16.1.14), a length of transmission line can transform a load Z_L to a range of possible complex values $Z(-l)$. To understand this range of values better, we can use the Smith chart (thanks to the genius of P.H. Smith who invented it in 1939 before the advent of the computer) [123]. The Smith chart is essentially a graphical calculator for solving transmission line problems. It has been used so much by microwave engineers during the early days that its use has left a strong imprint and legacy on these engineers: it also has become an indispensable visual and mental aids for understanding and solving microwave engineering problems. This chart occupies an important place in the hearts and minds of many microwave engineers; many of them can manipulate and predict the outcome of a design visually or mentally in their minds.

Equation (16.1.12) indicates that there is a unique, one-to-one map between the normalized impedance $Z_n(z) = Z(z)/Z_0$ and reflection coefficient $\Gamma(z)$. In the normalized impedance form where $Z_n = Z/Z_0$, from (16.1.11) and (16.1.12)

$$\Gamma = \frac{Z_n - 1}{Z_n + 1}, \qquad Z_n = \frac{1 + \Gamma}{1 - \Gamma} \tag{16.2.1}$$

Equations in (16.2.1) are known as bilinear transforms in complex variables [92]: It is a conformal map that maps circles to circles. Such a map is shown in Figure 16.4, where lines on the right-half of the complex Z_n plane are mapped to the circles on the complex Γ plane. Since straight lines on the complex Z_n plane are circles with infinite radii, they are mapped to circles on the complex Γ plane. The Smith chart shown on Figure 16.11 allows one to obtain the corresponding Γ given Z_n and vice versa as indicated in (16.2.1), but using a graphical calculator or the Smith chart. They can be done visually or mentally in many engineers' minds.

Notice that the imaginary axis on the complex Z_n plane maps to the circle of unit radius on the complex Γ plane. All points on the right-half plane are mapped to within the unit circle. The reason being that the right-half plane of the complex Z_n plane corresponds to passive impedances such that $R_n > 0$ that will absorb energy. Hence, by energy conservation, such an impedance load will have reflection coefficient with amplitude less than one, which are points within the unit circle as shown in Figure 16.4.

On the other hand, the left-half of the complex Z_n plane corresponds to impedances with negative resistances. These will be active elements that can generate energy, and hence, yielding $|\Gamma| > 1$; they correspond to points outside the unit circle on the complex Γ plane.

Another point to note is that points at infinity on the complex Z_n plane map to the point at $\Gamma = 1$ on the complex Γ plane, while the point zero on the complex Z_n plane maps to $\Gamma = -1$ on the complex Γ plane. These are the reflection coefficients of an open-circuit load and a short-circuit load, respectively. For a matched load, $Z_n = 1$ and $\Gamma = 0$ which maps to the zero point or the origin on the complex Γ plane implying no reflection. It looks like we are inundating you with a large amount of data here, and you may want to stop and contemplate on these points a bit!

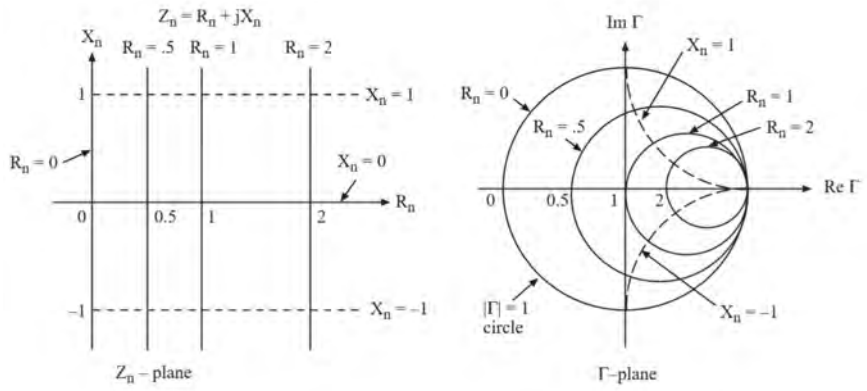


Figure 16.4: Bilinear map of the formulae $\Gamma = \frac{Z_n - 1}{Z_n + 1}$, and $Z_n = \frac{1+\Gamma}{1-\Gamma}$. It maps circles to circles. The chart on the right, called the Smith chart, allows the values of Z_n to be determined quickly once given Γ, and vice versa.

The Smith chart also allows one to quickly evaluate the expression[2]

$$\Gamma(-l) = \Gamma_L e^{-2j\beta l} \tag{16.2.2}$$

and its corresponding Z_n, not by using (16.2.1) via a calculator, but by using a graphical calculator—the Smith chart. Since $\beta = 2\pi/\lambda$, it is more convenient to write $\beta l = 2\pi l/\lambda$, and measure the length of the transmission line in terms of wavelength. To this end, the above becomes

$$\Gamma(-l) = \Gamma_L e^{-4j\pi(l/\lambda)} \tag{16.2.3}$$

In the above, the phase of the general reflection coefficient is changing in the unit l/λ, or that our yardstick is wavelength here.[3] For increasing l, one moves away from the load to the generator (or source). As l increases, the phase is decreasing because of the negative sign. So given a point for Γ_L on the Smith chart, one has a negative phase or a decreasing phase by rotating the point clockwise. Also, due to the $\exp(-4j\pi l/\lambda)$ dependence of the phase, when $l = \lambda/4$, the reflection coefficient acquires a phase factor of $\exp(-j\pi)$, which rotates it a half circle around the chart. And when $l = \lambda/2$, the reflection coefficient acquires a phase factor of $\exp(-j2\pi)$, which will rotate it full circle, or back to the original point.

Therefore, on the edge of the Smith chart, there are indications as to which direction one should rotate if one were to move toward the generator or toward the load. Again, there is a large amount of information here, but it is not rocket science. And it will be good if you can stop and contemplate a moment to wrap your head around these concepts.

For two points diametrically opposite to each other on the Smith chart, Γ changes sign, and it can be shown easily from (16.2.1) that the normalized impedances are reciprocal of each other.

[2]The factor of $2l$ in the exponent comes about because the wave has to travel a distance of $2l$ because the reflection coefficient is defined to be the ratio of the reflected wave to the incident wave at the location $z = -l$.

[3]Remember in electromagnetics, the yardstick is always the wavelength!

Hence, the Smith chart can also be used to find the reciprocal of a complex number quickly. This is because if

$$\Gamma = \frac{Z_n - 1}{Z_n + 1} \tag{16.2.4}$$

and we replace $Z_n \to 1/Z_n$, we see that $\Gamma \to -\Gamma$.

A full blown Smith chart is shown in Figure 16.11.

16.3 VSWR (Voltage Standing Wave Ratio)

From the previous section, one sees that the voltage and current are not constant on a transmission line. Therefore, one surmises that measuring the impedance of a device at microwave frequency is a tricky business. At low frequency, one can use an ohm meter with two wire probes to do such a measurement. But at microwave frequency, two pieces of wire become inductors, and two pieces of metal become capacitors. More sophisticated ways to measure the impedance are needed as described below.

Due to the interference between the forward traveling wave and the backward traveling wave, $V(z)$ is a function of position z on a terminated transmission line and it is given as

$$\begin{aligned} V(z) &= V_0 e^{-j\beta z} + V_0 e^{j\beta z}\Gamma_L \\ &= V_0 e^{-j\beta z}\left(1 + \Gamma_L e^{2j\beta z}\right) \\ &= V_0 e^{-j\beta z}\left(1 + \Gamma(z)\right) \end{aligned} \tag{16.3.1}$$

where we have used (16.1.6) for $\Gamma(z)$. Hence, $V(z)$ is not a constant but dependent on z, or that

$$|V(z)| = |V_0||1 + \Gamma(z)| \tag{16.3.2}$$

For lack of a better name, this is called the standing wave, even though it is not truly a standing wave: it is the interference pattern formed by two traveling waves in opposite directions.

In Figure 16.5, the relationship between variation of $1 + \Gamma(z)$ as z varies is shown.

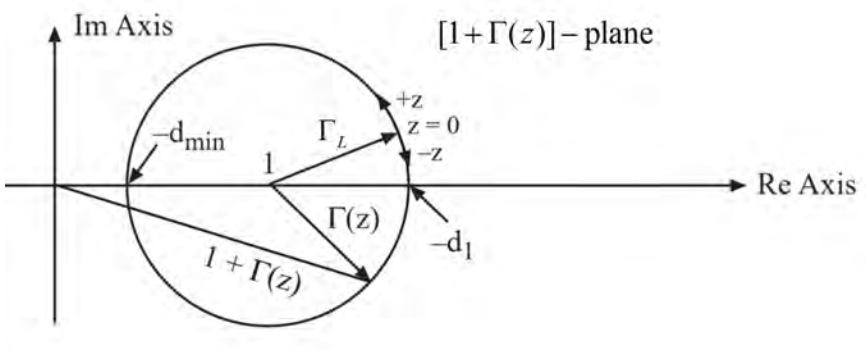


Figure 16.5: The voltage amplitude on a transmission line depends on $|V(z)|$, which is proportional to $|1+\Gamma(z)|$ per equation (16.3.2). This figure shows how $|1+\Gamma(z)|$ varies as z varies on a transmission line. Adding complex numbers on the complex plane is like adding vectors in 2D plane. $|1+\Gamma(z)|$ is the length of the "vector" that corresponds to the complex number $1+\Gamma(z)$.

Using the triangular inequality, one gets the lower and upper bounds or that

$$|V_0|(1-|\Gamma(z)|) \le |V(z)| \le |V_0|(1+|\Gamma(z)|) \tag{16.3.3}$$

But from (16.1.6) and that β is pure real for a lossless line, then $|\Gamma(z)| = |\Gamma_L|$; hence

$$V_{\min} = |V_0|(1-|\Gamma_L|) \le |V(z)| \le |V_0|(1+|\Gamma_L|) = V_{\max} \tag{16.3.4}$$

The voltage standing wave ratio, VSWR[4] is defined to be

$$\text{VSWR} = \frac{V_{\max}}{V_{\min}} = \frac{1+|\Gamma_L|}{1-|\Gamma_L|} \tag{16.3.5}$$

Conversely, one can invert the above formula to get

$$|\Gamma_L| = \frac{\text{VSWR}-1}{\text{VSWR}+1} \tag{16.3.6}$$

Hence, the knowledge of voltage standing wave pattern (VSWP) or its VSWR, as shown in Figure 16.6, yields the knowledge of $|\Gamma_L|$, the amplitude of Γ_L. Notice that the relations between VSWR and $|\Gamma_L|$ are homomorphic to those between Z_n and Γ. Therefore, the Smith chart can also be used to evaluate the above equations.

[4]This word has no vowel, and you pronounce it like a word in Hebrew.

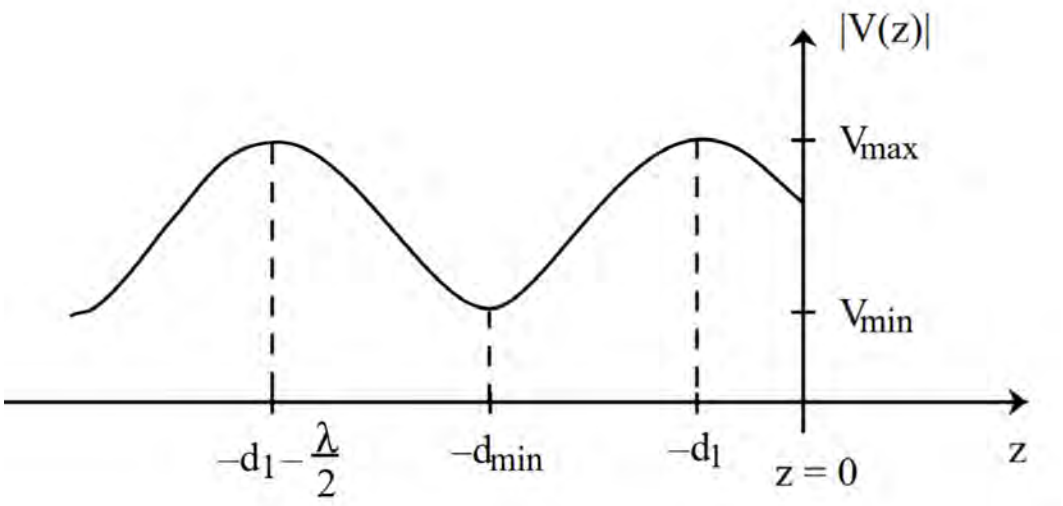


Figure 16.6: The voltage standing wave pattern (VSWP) as a function of z on a load-terminated transmission line.

The phase of Γ_L can also be determined from the measurement of the VSWP. The location of Γ_L in the complex Γ plane in Figure 16.5 is determined by the phase of Γ_L. Hence, the value of d_1 in Figure 16.5 is determined by the phase of Γ_L as well. The length of the transmission line waveguide needed to cancel the original phase of Γ_L to bring the voltage standing wave pattern to a maximum value at $z = -d_1$ is shown in Figure 16.6. Thus, d_1 is the value where the following equation is satisfied:

$$|\Gamma_L| e^{j\phi_L} e^{-4\pi j(d_1/\lambda)} = |\Gamma_L| \tag{16.3.7}$$

Therefore, by measuring the voltage standing wave pattern, one deduces both the amplitude and phase of Γ_L. From the complex value Γ_L, one can determine Z_L, the load impedance from the Smith chart.

Figure 16.7: A slotted-line equipment which consists of a coaxial waveguide with a slot opening at the top to allow the measurement of the field strength, and hence, the voltage standing wave pattern in the waveguide (courtesy of Microwave101.com).

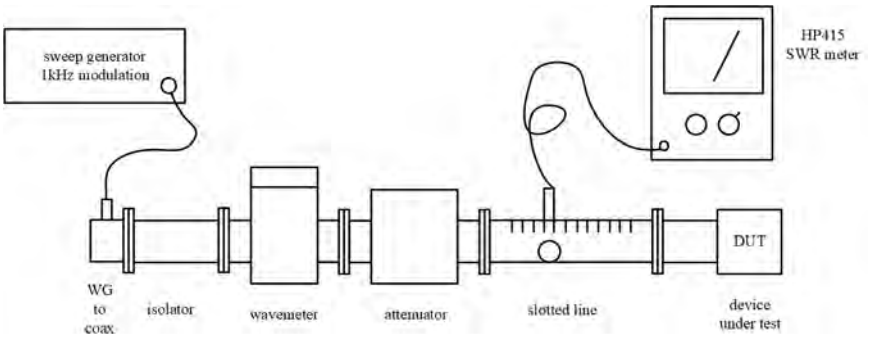


Figure 16.8: An experimental setup for a slotted line measurement (courtesy of Pozar and Knapp, U. Mass [124]).

In the old days, the voltage standing wave pattern was measured by a slotted-line equipment which consists of a coaxial waveguide with a slot opening as shown in Figure 16.7. A field probe can be inserted into the slotted line to determine the strength of the electric field inside the coax waveguide. A typical experimental setup for a slotted line measurement is shown in Figure 16.8. A generator source, with low frequency modulation, feeds microwave energy into the coaxial waveguide. The isolator, allowing only the unidirectional propagation of microwave energy, protects the generator. The attenuator protects the slotted line equipment. The wavemeter is an adjustable resonant cavity. When the wavemeter is tuned to the frequency of the microwave, it siphons off some energy from the source, giving rise to a dip in the signal of the SWR meter (a shorthand for voltage-standing-wave-ratio meter). Hence, the wavemeter measures the frequency of the microwave.

The slotted line probe is usually connected to a square law detector with a rectifier that converts the microwave signal to a low-frequency signal. In this manner, the amplitude of the voltage in the slotted line can be measured with some low-frequency equipment, such as the SWR meter.

Low-frequency equipment is a lot cheaper to make and maintain. That is also the reason why the source is modulated with a low-frequency signal. At low frequencies, circuit theory prevails, making engineering and design a lot simpler.

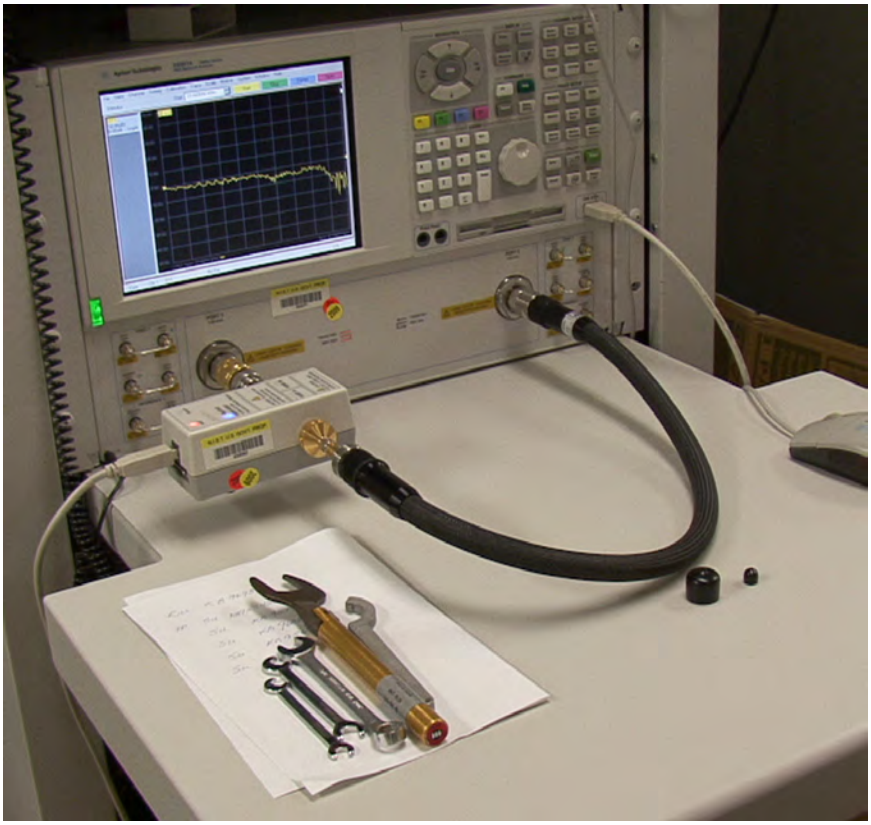


Figure 16.9: The microwave metrology technologies have progressed by leaps and bounds. I had to use a slotted line to measure the impedance of a device when I was a student. Now, a vector (measures both phase and amplitude) automated network analyzer is used and shown. It makes the measurements of microwave parameters a lot easier (courtesy of NIST).

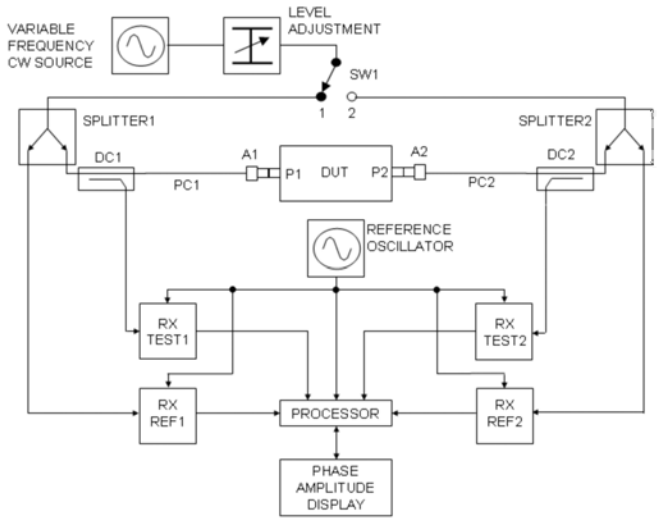


Figure 16.10: A schematic of the automated network analyzer. You can get the definitions of the acronyms from [125] (courtesy of Wikipedia).

The above describes how the impedance of the device-under-test (DUT) can be measured at microwave frequencies. Nowadays, automated network analyzers (ANA) make these measurements a lot simpler in a microwave laboratory. A picture of an ANA is shown in Figure 16.9. A schematic of how it works is shown in Figure 16.10. More resource on microwave measurements can be found on the web, such as in [126].

Notice that the above is based on the interference of the two traveling wave on a terminated transmission line. Such interference experiments are increasingly difficult in optical frequencies because of the much shorter wavelengths. As such, many experiments are easier to perform at microwave frequencies rather than at optical frequencies.

Many electromagnetics technologies were first developed at microwave frequency, and later developed at optical frequency. Examples are phase imaging, optical coherence tomography, and beam steering with phase array sources. Another example is that quantum information and quantum computing can be done at optical frequency, but the recent trend is to use artificial atoms working at microwave frequencies. Engineering with longer wavelength and larger component is easier; and hence, microwave engineering is employed. For instance, the recent Sycamore quantum computer made by Google and the Chinese ZuChongZhi quantum computer uses hoards of microwave-engineering concepts [127, 128].

Another new frontier in the electromagnetic spectrum is in the terahertz range. Due to the dearth of sources in the terahertz range, and the added difficulty in having to engineer with smaller components, this is an exciting and a largely untapped frontier in electromagnetic technology.

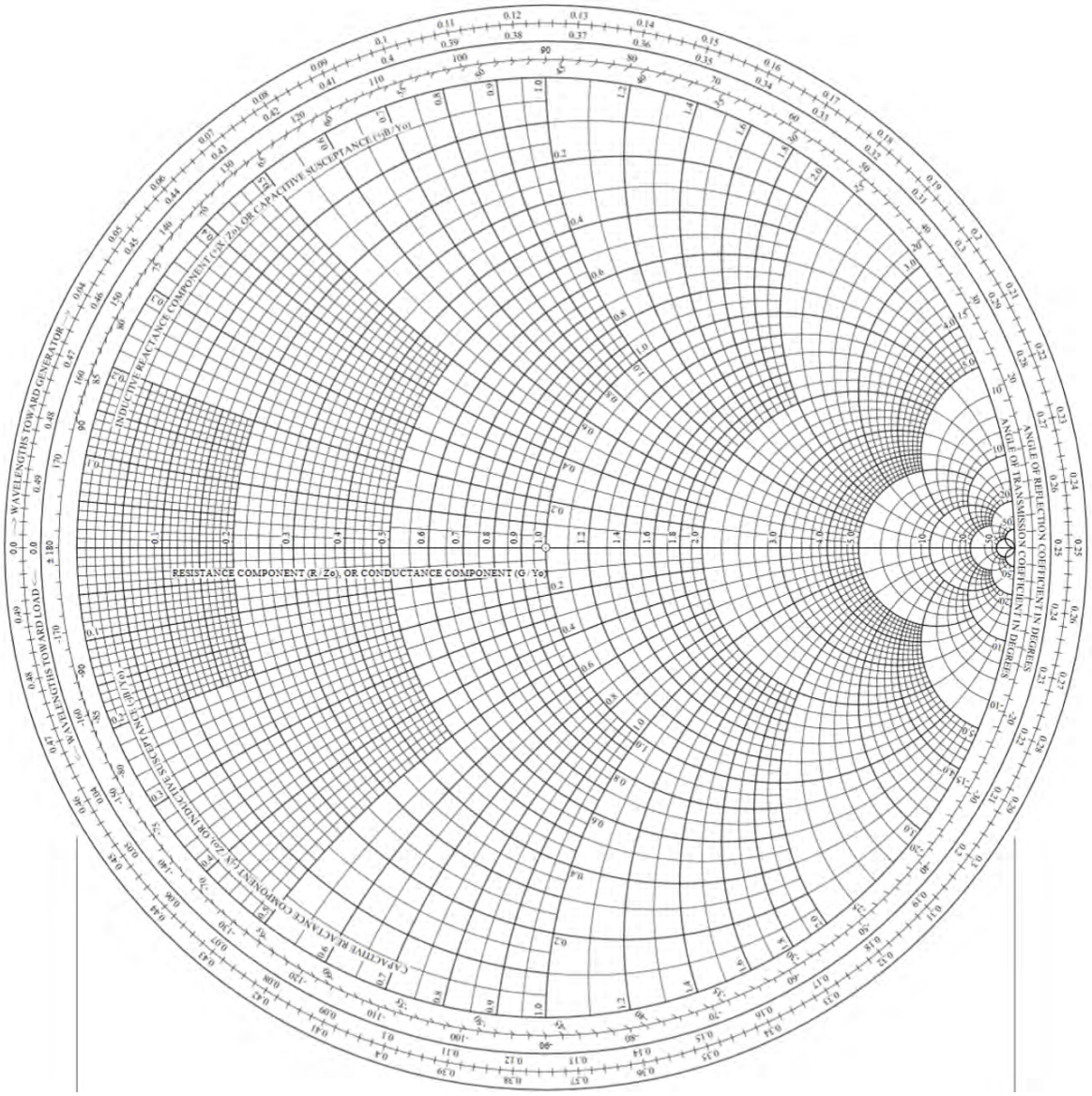


Figure 16.11: The Smith chart in its full glory. It was invented in 1939 before the age of digital computers, but it still allows microwave engineers to do mental estimations, gymnastics, and rough predictions with it. Because of its simplicity, it has an indelible influence on microwave engineering.

Exercises for Lecture 16

Problem 16-1: Look at Lecture 8 of the notes from ECE350x: https://engineering.purdue.edu/wcchew/ece350/ee350-08.pdf

(i) Walk yourself through the example, for part (a) change the load to

$$Z_L = 20 + j30\ \Omega$$

and find the new answer.

(ii) For part (b), change

$$d_{\min} = 3\lambda/16$$

and find the new answer.

Chapter 17

Multi-Junction Transmission Lines, Duality Principle

A simple extension of the transmission line theory of the previous lecture is to the case when transmission lines of different characteristic impedances and wave velocities are concatenated together. Myriads of devices can be designed using such admixture of transmission lines, transistors, and diodes. Microwave engineering is a vibrant field because much of it can be described by transmission line theory, a poor-man's Maxwell's equations. The wisdom of our predecessors is that the simpler the concepts are, they easier they can be engineered. The folklore is that when Maxwell completed his treatise in electricity and magnetism [112], few could read beyond the first fifty pages of his tome. It is through decades of regurgitation, distillation and simplification that we can now teach this knowledge to undergraduate students. Also, much of microwave components are of thumb size or hand size, making the engineering of their wave-physics components easier compared to radio waves and optics.

Therefore, due to the symmetry of Maxwell's equations between the electric field and the magnetic field, once a set of solutions has been found for Maxwell's equations, new solutions can be found by symmetry arguments. This is known as duality principle in electromagnetics. Recently, the use of symmetry of Maxwell's equations has given rise to the field of metamaterials. This field holds promise for new materials that can offer new physical phenomena [129, 130].

17.1 Multi-Junction Transmission Lines

The real world is usually more complex than the world of our textbooks. However, we need to distill problems in the real world into simpler problems so that we can explain them with our textbook examples. Figure 17.1 shows many real world technologies, but they can be approximated with transmission line models as shall be seen.

Figure 17.1: Different kinds of waveguides operating in different frequencies in power lines, RF circuits, microwave circuits, and optical fibers. Their salient physics or features can be captured or approximated by transmission lines (courtesy of Owen Casha).

An area where multi-junction transmission lines play an important role is in the microwave integrated circuits (MIC) and the monolithic microwave integrated circuits (MMIC). An MMIC circuit is shown in Figure 17.2. Many of the components can be approximated by multi-junction transmission lines. Thus they are clear motivation for studying this topic.

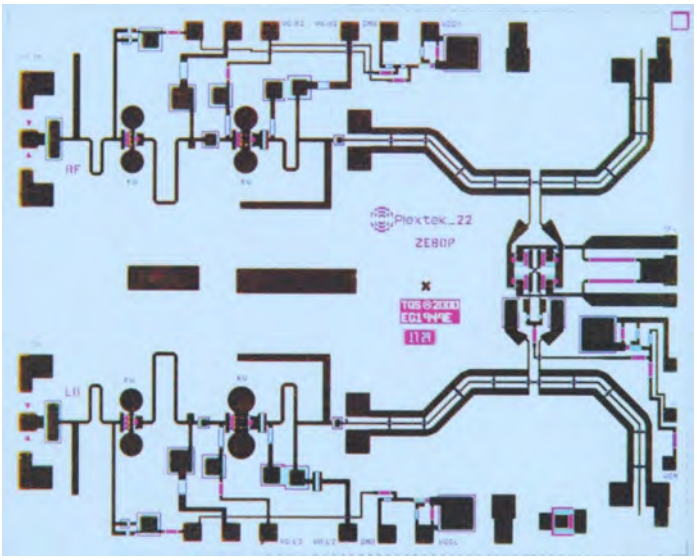


Figure 17.2: A generic GaAs monolithic microwave integrated circuit (MMIC). They are the motivation for studying multi-junction transmission lines (courtesy of Wikipedia).

By concatenating sections of transmission lines of different characteristic impedances, a large variety of devices such as resonators, filters, radiators, and matching networks can be formed. We will start with a single junction transmission line first. Good references for such a problem are the books by Collin [131] and Pozar [132], but much of the treatment here is not found in any textbooks.

17.1.1 Single-Junction Transmission Lines

Consider two transmission lines connected at a single junction as shown in Figure 17.3. For simplicity, we assume that the transmission line to the right is infinitely long so that a right-traveling wave is not reflected; namely, there is no reflected wave. And we assume that the two transmission lines have different characteristic impedances, Z_{01} and Z_{02}. As such, due to the impedance mismatch at junction 1, a right-traveling wave in line 1 is reflected at junction 1.

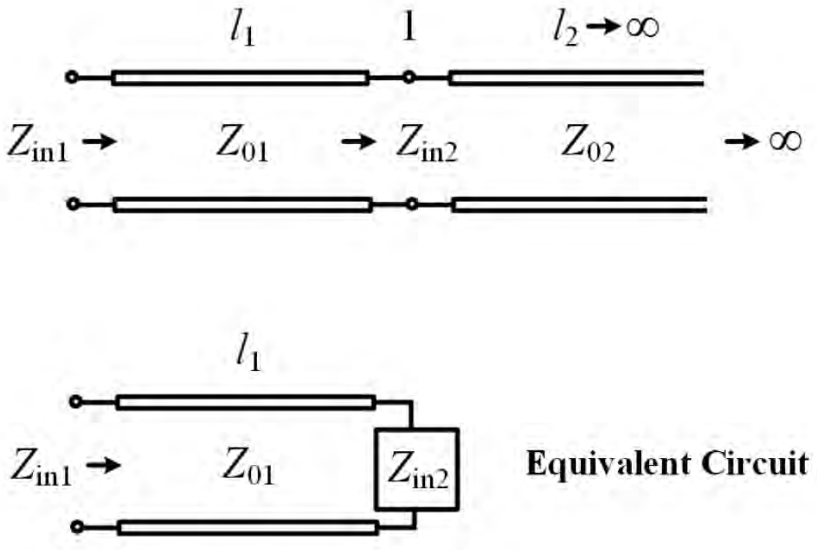


Figure 17.3: A single junction transmission line can be modeled by an equivalent transmission line terminated in a load Z_{in2}.

The impedance of the transmission line at junction 1 looking to the right (see Figure 17.3),using the formula from previously derived,[1] is

$$Z_{in2} = Z_{02}\frac{1+\Gamma_{L,\infty}e^{-2j\beta_2 l_2}}{1-\Gamma_{L,\infty}e^{-2j\beta_2 l_2}} = Z_{02} \tag{17.1.1}$$

[1]We should always remember that the relations between the reflection coefficient Γ and the normalized impedance Z_n are $\Gamma = \frac{Z_n-1}{Z_n+1}$ and $Z_n = \frac{1+\Gamma}{1-\Gamma}$.

We have let $\Gamma_{L,\infty} = 0$ since no reflected wave exists on the last section. Thus the above is just Z_{02}. As a result, transmission line 1 sees a load of $Z_L = Z_{in2} = Z_{02}$ hooked to its end. The equivalent circuit is shown in Figure 17.3 as well. Hence, we deduce that the reflection coefficient at junction 1 between line 1 and line 2, using the knowledge from the previous lecture, is Γ_{12}, and is given by

$$\Gamma_{12} = \frac{Z_L - Z_{01}}{Z_L + Z_{01}} = \frac{Z_{in2} - Z_{01}}{Z_{in2} + Z_{01}} = \frac{Z_{02} - Z_{01}}{Z_{02} + Z_{01}} \tag{17.1.2}$$

where we have assumed that $Z_L = Z_{in2} = Z_{02}$.

17.1.2 Two-Junction Transmission Lines—Composite Reflection Coefficient

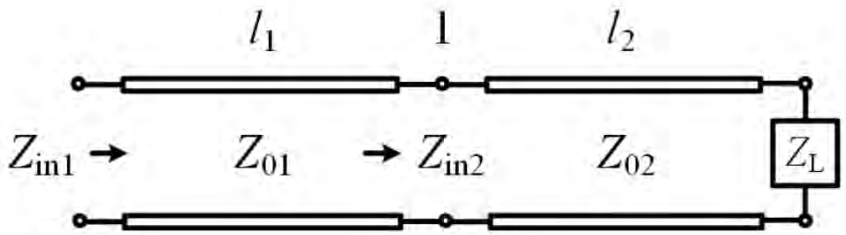


Figure 17.4: A two-junction transmission line can be modeled by a single-junction transmission line with a load. The last section (third section) is infinitely long and can be replaced with a load Z_L at the far (load) end of the second line. But it can be reduced to the equivalent circuit shown in the bottom of Figure 17.3.

Now, we look at the two-junction case. To this end, we first look at when line 2 is terminated by a load Z_L at its load end as shown in Figure 17.4. Then, using the formula derived in the previous lecture,

$$Z_{in2} = Z_{02}\frac{1 + \Gamma(-l_2)}{1 - \Gamma(-l_2)} = Z_{02}\frac{1 + \Gamma_{L2}e^{-2j\beta_2 l_2}}{1 - \Gamma_{L2}e^{-2j\beta_2 l_2}} \tag{17.1.3}$$

where we have used the fact that $\Gamma(-l_2) = \Gamma_{L2}e^{-2j\beta_2 l_2}$. It is to be noted that here, using knowledge from the previous lecture, that the reflection coefficient at the load end of line 2 is

$$\Gamma_{L2} = \frac{Z_L - Z_{02}}{Z_L + Z_{02}} \tag{17.1.4}$$

Now, line 1 sees a load of Z_{in2} hooked at its end. The equivalent circuit is the same as that shown in Figure 17.3. The composite reflection coefficient at junction 1, which includes all the

reflected waves from its right, is now[2]

$$\tilde{\Gamma}_{12} = \frac{Z_{in2} - Z_{01}}{Z_{in2} + Z_{01}} \tag{17.1.5}$$

To elaborate further, we substitute (17.1.3) into (17.1.5), to arrive at

$$\tilde{\Gamma}_{12} = \frac{Z_{02}(\frac{1+\Gamma}{1-\Gamma}) - Z_{01}}{Z_{02}(\frac{1+\Gamma}{1-\Gamma}) + Z_{01}} \tag{17.1.6}$$

where $\Gamma = \Gamma_{L2} e^{-2j\beta_2 l_2}$. The above can be rearranged to give

$$\tilde{\Gamma}_{12} = \frac{Z_{02}(1+\Gamma) - Z_{01}(1-\Gamma)}{Z_{02}(1+\Gamma) + Z_{01}(1-\Gamma)} \tag{17.1.7}$$

Finally, by further rearranging terms, after a somewhat tedious algebra, it can be shown that the above becomes

$$\tilde{\Gamma}_{12} = \frac{\Gamma_{12} + \Gamma}{1 + \Gamma_{12}\Gamma} = \frac{\Gamma_{12} + \Gamma_{L2} e^{-2j\beta_2 l_2}}{1 + \Gamma_{12}\Gamma_{L2} e^{-2j\beta_2 l_2}} \tag{17.1.8}$$

where Γ_{12}, the local reflection coefficient at the junction between line 1 and line 2, is given by (17.1.2), and $\Gamma = \Gamma_{L2} e^{-2j\beta_2 l_2}$ is the local reflection coefficient[3] at $z = -l_2$ due to the load Z_L. In other words, referring to Figure 17.4, with knowledge of the local reflection coefficient Γ_{12} at the junction 1 in accordance to (17.1.2) and the reflection coefficient Γ_{L2} at the load end according to (17.1.4), one can use the above formula to find the composite reflection coefficient $\tilde{\Gamma}_{12}$ at junction 1. The composite reflection coefficient includes all reflections of the wave to the right of the junction 1, while the local reflection coefficient Γ_{12} is due only to the impedance difference between line 1 and line 2 at junction 1. It is the reflected wave "locally" off the (1,2) junction.

17.1.3 Recursive Formula for Composite Reflection Coefficient

Equation (17.1.8) is a *powerful formula* for multi-junction transmission lines. Imagine now that we add another section of transmission line as shown in Figure 17.5. We can use the aforementioned method to first find $\tilde{\Gamma}_{23}$, the composite reflection coefficient at junction 2. Using formula (17.1.8), it is given by

$$\tilde{\Gamma}_{23} = \frac{\Gamma_{23} + \Gamma_{L3} e^{-2j\beta_3 l_3}}{1 + \Gamma_{23}\Gamma_{L3} e^{-2j\beta_3 l_3}} \tag{17.1.9}$$

where Γ_{L3} is the load reflection coefficient due to the load Z_L hooked to the end of transmission line 3 as shown in Figure 17.5. Here, it is given as

$$\Gamma_{L3} = \frac{Z_L - Z_{03}}{Z_L + Z_{03}} \tag{17.1.10}$$

[2] I used to call this reflection coefficient "generalized reflection coefficient" [108], but this name seems to have been hijacked by some authors in the community for what we call "the general reflection coefficient" in this course.

[3] We will use the term "local reflection coefficient" at location z to mean the ratio between the amplitudes of the left-traveling wave and the right-traveling wave on a transmission line at z due to the discontinuity in the impedance at the (1,2) junction.

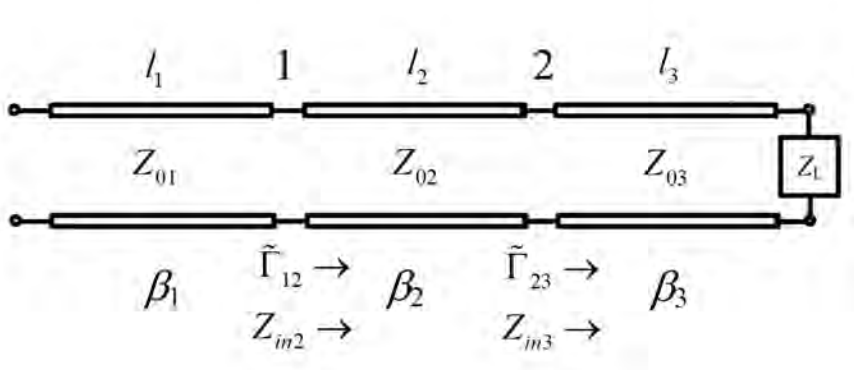


Figure 17.5: A two-junction transmission line with a load Z_L at the far end. The input impedance looking in from the far left can be found recursively using the formula (17.1.11) and (17.1.12).

Given the knowledge of $\tilde{\Gamma}_{23}$, we can use (17.1.8) again to find the new $\tilde{\Gamma}_{12}$ at junction 1. It is now

$$\tilde{\Gamma}_{12} = \frac{\Gamma_{12} + \tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}}{1 + \Gamma_{12}\tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}} \tag{17.1.11}$$

The equivalent circuit is again that shown in Figure 17.3. Therefore, we can use (17.1.11) recursively to find the composite reflection coefficient for a multi-junction transmission line. Once the reflection coefficient is known, the impedance at that location can also be found.[4] For instance, at junction 1, the impedance is now given by

$$Z_{in2} = Z_{01} \frac{1 + \tilde{\Gamma}_{12}}{1 - \tilde{\Gamma}_{12}} \tag{17.1.12}$$

instead of (17.1.3). In the above, Z_{01} is used because the composite reflection coefficient $\tilde{\Gamma}_{12}$ is the total reflection coefficient for an incident wave from transmission line 1 that is sent toward the junction 1. Previously, Z_{02} was used in (17.1.3) because the reflection coefficients in that equation was for an incident wave sent from transmission line 2.

If the incident wave were to have come from line 2, then one can write Z_{in2} as

$$Z_{in2} = Z_{02} \frac{1 + \tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}}{1 - \tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}} \tag{17.1.13}$$

With some algebraic manipulation, it can be shown that (17.1.12) and (17.1.13) are identical. Therefore, it is important to envision in our mind an incident wave and a reflected wave, and on which line these waves are traveling. But (17.1.12) is closer to an experimental scenario where one

[4] We have found from the previous lecture that there is a one-to-one map between reflection coefficient and the normalized impedance at the location z.

measures the reflection coefficient by sending a wave from line 1 with no knowledge of what is to the right of junction 1.

Transmission lines can be made easily in microwave integrated circuit (MIC) by etching or milling. A picture of a microstrip line waveguide or transmission line is shown in Figure 17.6.

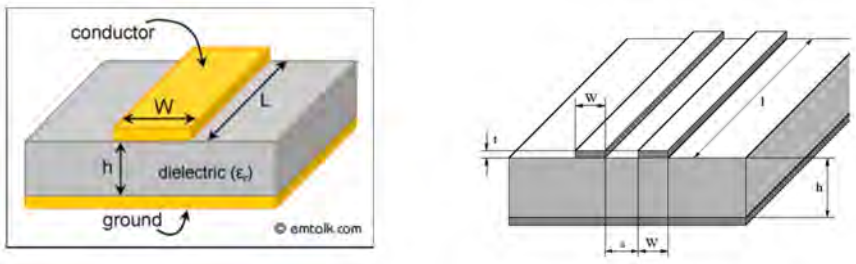


Figure 17.6: (Left) Schematic of a microstrip line with the signal line above, and a ground plane below. (Right) A strip line with each strip carrying currents of opposite polarity. A ground plane is not needed in the second case. Unlike the coaxial transmission line, there is no closed form solution for these geometries. Thanks to efficient computational electromagnetics (CEM) algorithms, numerical solutions to these problems exist (left figure, courtesy of emtalk.com, right figure, courtesy of qucs.sourceforge.net).

17.1.4 Stray Capacitance and Inductance

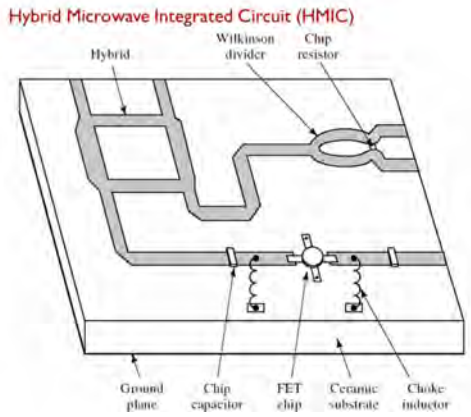


Figure 17.7: A general microwave integrated circuit with different kinds of elements (courtesy of slideplayer.com).

The junction between two transmission lines is not as simple as we have assumed. In the real world, or in MIC, the waveguide junction has discontinuities in line width, or shape. This can give

rise to excess charge accumulation or constricted current flow. Excess charge gives rise to excess electric field which corresponds to excess electric stored energy. This can be modeled by a stray or parasitic capacitance C_s as shown in Figure 17.8.

Alternatively, there could be constricted current flow that gives rise to stronger magnetic field.[5] Excess magnetic field compared to normal gives rise to excess magnetic stored energy. This can be modeled by stray or parasitic inductances L_{s1} and L_{s2}. Hence, a junction can be approximated by a circuit model as shown in Figure 17.8 to account for these effects. The Smith chart or the method we have outlined above can still be used to solve for the input impedances of a transmission circuit when these parasitic circuit elements are added.

Notice that when the frequency is zero or low, these stray capacitances and inductances are negligible. We retrieve the simple junction model. But since their impedance and admittance are $j\omega L_s$ and $j\omega C_s$, respectively, they are non-negligible and are instrumental in modeling high frequency circuits.

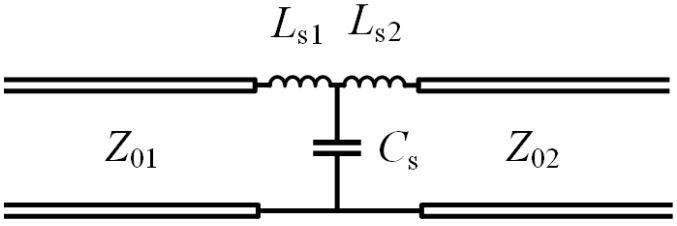


Figure 17.8: A junction between two microstrip lines can be modeled with a stray junction capacitance and stray inductances. The capacitance is used to account for excess charges at the junction, while the inductances model the excess current at the junction. They are important as the frequency increases.

[5]The magnetic field around a thinner wire is stronger than that of a thicker wire.

17.1.5 Multi-Port Network

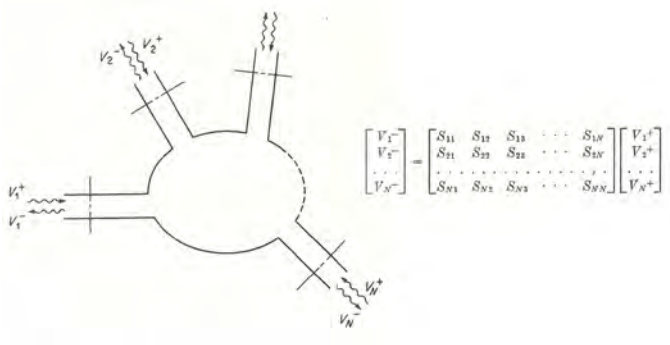


Figure 17.9: A multi-port network can be characterized by its scattering matrix and the scattering parameters. Once known, microwave circuits can be concatenated together to form larger complex circuits. If the scattering parameters are known over a broad bandwidth, the behavior of the circuit in the time domain can also be obtained by Fourier transform (courtesy of Collin [101]).

A section of transmission line can be thought of as a two-port network as in circuit theory, but with the difference that the inputs are the incident waves at each of these port, and the outputs are the reflected waves at the ports. This concept can be generalized to a multi-port network with N ports easily. The inputs are the incident voltage waves (on a transmission line), and the outputs are the reflected voltage waves at each of the ports. Since the system is linear, in general, the inputs and outputs are linearly related by a scattering matrix, or

$$\mathbf{V}^- = \overline{\mathbf{S}} \cdot \mathbf{V}^+ \tag{17.1.14}$$

The scattering matrix is a very useful and important microwave engineering concept. It encapulates or characterizes the properties of a complex microwave circuit with numbers called the scattering parameters. The scattering parameters can be measured or calculated. Once the scattering matrix of a microwave circuit is known, it can be concatenated with other microwave circuits similarly characterized.

17.2 Duality Principle

Duality principle exploits the inherent symmetry of Maxwell's equations. Once a set of **E** and **H** fields have been found to solve Maxwell's equations for a certain geometry, another set for a similar geometry can be found by invoking this principle. Maxwell's equations in the frequency domain,

including the fictitious magnetic sources, are

$$\nabla \times \mathbf{E}(\mathbf{r},\omega) = -j\omega\mathbf{B}(\mathbf{r},\omega) - \mathbf{M}(\mathbf{r},\omega) \tag{17.2.1}$$

$$\nabla \times \mathbf{H}(\mathbf{r},\omega) = j\omega\mathbf{D}(\mathbf{r},\omega) + \mathbf{J}(\mathbf{r},\omega) \tag{17.2.2}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r},\omega) = \varrho_m(\mathbf{r},\omega) \tag{17.2.3}$$

$$\nabla \cdot \mathbf{D}(\mathbf{r},\omega) = \varrho(\mathbf{r},\omega) \tag{17.2.4}$$

One way to make Maxwell's equations invariant is to perform the following substitutions (or transformation).

$$\mathbf{E} \to \mathbf{H}, \quad \mathbf{H} \to -\mathbf{E}, \quad \mathbf{D} \to \mathbf{B}, \quad \mathbf{B} \to -\mathbf{D} \tag{17.2.5}$$

$$\mathbf{M} \to -\mathbf{J}, \quad \mathbf{J} \to \mathbf{M}, \quad \varrho_m \to -\varrho, \quad \varrho \to \varrho_m \tag{17.2.6}$$

The above swaps retain the right-hand rule for plane waves. When material media is included, such that $\mathbf{D} = \overline{\varepsilon} \cdot \mathbf{E}$, $\mathbf{B} = \overline{\mu} \cdot \mathbf{H}$, for anisotropic media, Maxwell's equations become

$$\nabla \times \mathbf{E} = -j\omega\overline{\mu} \cdot \mathbf{H} - \mathbf{M} \tag{17.2.7}$$

$$\nabla \times \mathbf{H} = j\omega\overline{\varepsilon} \cdot \mathbf{E} + \mathbf{J} \tag{17.2.8}$$

$$\nabla \cdot \overline{\mu} \cdot \mathbf{H} = \varrho_m \tag{17.2.9}$$

$$\nabla \cdot \overline{\varepsilon} \cdot \mathbf{E} = \varrho \tag{17.2.10}$$

In addition to the above swaps, one needs further to swap the material parameters, namely,

$$\overline{\mu} \to \overline{\varepsilon}, \quad \overline{\varepsilon} \to \overline{\mu} \tag{17.2.11}$$

17.2.1 Unusual Swaps[6]

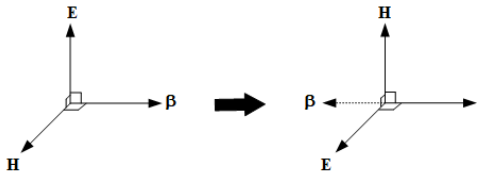


Figure 17.10: Unusual swap expounded in (17.2.12), though leaving Maxwell's equations unchanged, seems to disobey the right-hand rule for $\mathbf{E}$, $\mathbf{H}$, and $\boldsymbol{\beta}$. But one can change the sign of β as it is not unique to obey the right-hand rule again.

[6]This section can be skipped on first reading.

There are other swaps where seemingly the right-hand rule is not preserved (see Figure 17.10), e.g.,

$$\mathbf{E} \to \mathbf{H},\ \mathbf{H} \to \mathbf{E},\ \mathbf{M} \to -\mathbf{J},\ \mathbf{J} \to -\mathbf{M}, \tag{17.2.12}$$

$$\varrho_m \to -\varrho,\ \varrho \to -\varrho_m,\ \overline{\boldsymbol{\mu}} \to -\overline{\boldsymbol{\varepsilon}},\ \overline{\boldsymbol{\varepsilon}} \to -\overline{\boldsymbol{\mu}} \tag{17.2.13}$$

The above swaps will leave Maxwell's equations invariant, but when applied to a plane wave, the right-hand rule seems violated.

The deeper reason is that solutions to Maxwell's equations are not unique, since there is a time-forward as well as a time-reverse solution. In the frequency domain, this shows up in the choice of the sign of the $\mathbf{k}$ vector where in a plane wave $k = \pm\omega\sqrt{\mu\varepsilon}$. When one does a swap of $\mu \to -\varepsilon$ and $\varepsilon \to -\mu$, k is still indeterminate, and one can always choose a root where the right-hand rule is retained.

17.2.2 Left-Handed Materials and Double Negative Materials

The above unusual swap reminds us of the double-negative (DNG) materials or left-handed materials (LHM) that have inspired some recent works in metamaterials in electromagnetics [129, 130]. Assuming a simple source-free homogeneous-medium case where we have let $\mu \to -\mu$ and $\varepsilon \to -\varepsilon$ to arrive at

$$\nabla \times \mathbf{E} = j\omega\mu\mathbf{H} \tag{17.2.14}$$

$$\nabla \times \mathbf{H} = -j\omega\varepsilon\mathbf{E} \tag{17.2.15}$$

If we further assume a plane-wave solution in the above and let the space dependence of the solution to be $\exp(-j\boldsymbol{\beta}\cdot\mathbf{r})$, such a plane wave solution obeys the left-hand rule rather than the right-hand rule. Hence, such a material, first proposed by Veselago [129], and later promulgated by Pendri [26], had been a hot topic of research. Since $\mathbf{E}(\mathbf{r},t) = \Re e\left\{\mathbf{E}(\mathbf{r},\omega)e^{-j\omega t}\right\}$, the above equations can also be obtained by letting $t \to -t$, or by letting $j \to -j$. The above can be thought of as a left-handed solution traveling forward in time, or a right-handed solution traveling backward in time.

17.3 Fictitious Magnetic Currents

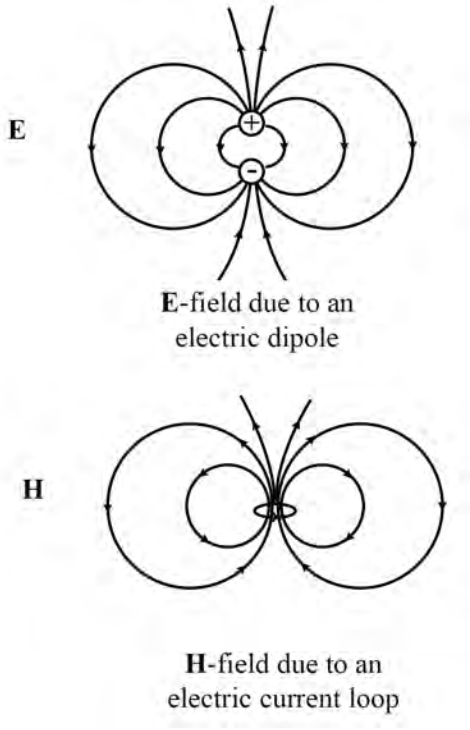


Figure 17.11: Sketches of the electric field **E** due to an electric dipole and the magnetic field **H** due to a electric current loop. The **E** and **H** fields have the same pattern, and can be described by the same formula. Hence, the magnetic field of a current loop resembles that of a magnetic dipole. As such, a current loop is a good mimicry of a magnetic dipole.

Even though magnetic charges or monopoles do not exist, magnetic dipoles do. For instance, a magnet can be regarded as a magnetic dipole. Also, it is believed that electrons have spins, and these spins make electrons behave like tiny magnetic dipoles in the presence of a magnetic field.

Also if we form electric current into a loop, it produces a magnetic field that looks like the electric field of an electric dipole. This resembles a magnetic dipole field. Hence, a magnetic dipole can be "mimicked" using a small electric current loop (see Figure 17.11). The magnetic field external to the current loop is essentially that of a magnetic dipole. Because of these similarities, it is common to introduce fictitious magnetic charges and magnetic currents into Maxwell's equations. One can think that these magnetic charges always occur in pair. Thus, they do not contradict the absence of magnetic monopole.

The electric current loops can be connected in series to make a toroidal antenna as shown

in Figure 17.12. The toroidal antenna is used to drive a current in an electric dipole. Notice that the toroidal antenna ressembles the primary winding of a transformer circuit. In essence, the toroidal loops, which mimic a magnetic current loop, produces an electric field that will drive current through the cylinder forming a long electric dipole. This is dual to the fact that an electric current loop produces a magnetic field. This is the working principle behind the measurement-while-drilling tool in the oil industry [133]. The entire drill stem inside a well bore for well logging can be used as an antenna. It also serves as a Goubau line [134].

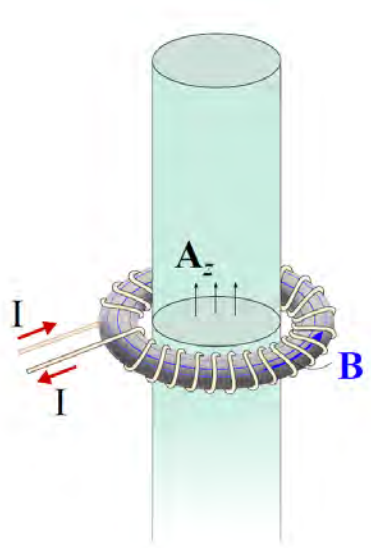


Figure 17.12: A toroidal antenna used to drive an electric current through a conducting cylinder of a dipole. It works similarly to a transformer: one can think of them as the primary and secondary turns (windings) of a transformer. Alternatively, one can think of the sequence of small electric current loops as forming a larger magnetic current loop. This larger magnetic current loop produces an electric field that drives an electric current through the conducting cylinder (courtesy of Q. S. Liu [135]).

Exercises for Lecture 17

Problem 17-1: The multi-section (or junction) transmission line is as shown in the figure below.

(i) Use the composite reflection coefficient derived in class, find $\tilde{\Gamma}_{23}$ and Z_{in3}.

(ii) Then find $\tilde{\Gamma}_{12}$ and Z_{in2}.

(iii) What is the value of Z_{02} you can choose to have zero reflection at Junction 1? (Note: This problem can also be solved using the graphical calculator, the Smith chart, but the closed form formulas allow one to calculate the reflection coefficients and the impedances exactly. Part (iii) of this problem is that of a quarter wave transformer matching which can be found in many textbooks.)

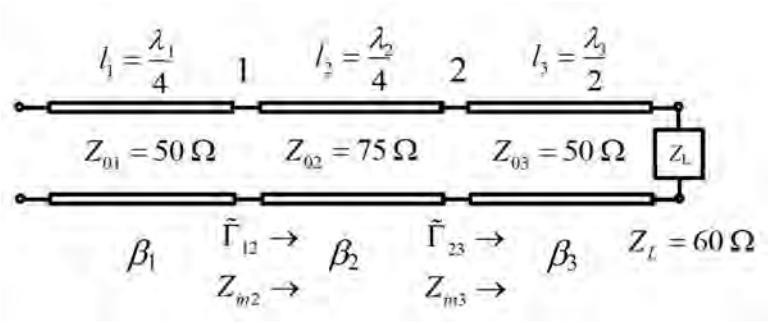


Figure 17.13: A transmission-line matching using quarter-wave transformer.

Chapter 18

Reflection, Transmission, and Interesting Physical Phenomena

We have seen the derivation of a reflection coefficient in a transmission line that relates the amplitude of the reflected wave to that of the incident wave. By doing so, we have used a simplified form of Maxwell's equations, the telegrapher's equations which are equations in one dimension. Here, we will solve Maxwell's equations in its full glory, but in order to do so, we will look at a very simple problem of plane wave reflection and transmission at a single plane interface.

This study will result in the Fresnel reflection and transmission coefficients, and embedded in them are interesting physical phenomena as shall be shown. We shall elucidate the physics of these interesting phenomena as well. This is a rare example of a simple closed-form solution to Maxwell's equations that offers scientists physical insight to the vast number of electromagnetic solutions. [1] Even though this is only the tip of the iceberg, it offers physical insight into the interaction of wave field with a simple medium or geometry.

(Much of the contents of this lecture can be found in Kong, and also the ECE 350X lecture notes. They can be found in many textbooks, even though the notations can be slightly different [53, 121, 57, 85, 33, 68, 34, 36, 90, 47].)

18.1 Reflection and Transmission—Single Interface Case

We will derive the plane-wave reflection and transmission coefficients in closed-form for the single interface case between two dielectric media. These reflection and transmission coefficients are the Fresnel reflection and transmission coefficients because they were first derived by Austin-Jean Fresnel (1788-1827).[2]

[1] One notable point is that glass spectacles were made in China as early as the Eastern Han dynasty (AD 25-220) [136] and law of refraction was known in the Islamic world around AD 900 [137].

[2] Note that he lived before the completion of Maxwell's equations in 1865. But when Fresnel derived the these coefficients in 1823, they were based on the elastic theory of light; and hence, the formulas are not exactly the same as what we are going to derive (see Born and Wolf, Principles of Optics, p. 40 [61]).

The single plane interface, plane wave reflection and transmission problem, with complicated mathematics, is "homomorphic" to the transmission line problem. The complexity comes because we have to keep track of the 3D polarizations of the electromagnetic fields in this case, as well as finding a solution in 3D space. We shall learn later that the mathematical "homomorphism" can be used to exploit the simplicity of transmission line theory in seeking the solutions to the multiple dielectric interface problems.

18.1.1 TE Polarization (Perpendicular or E Polarization)[3]

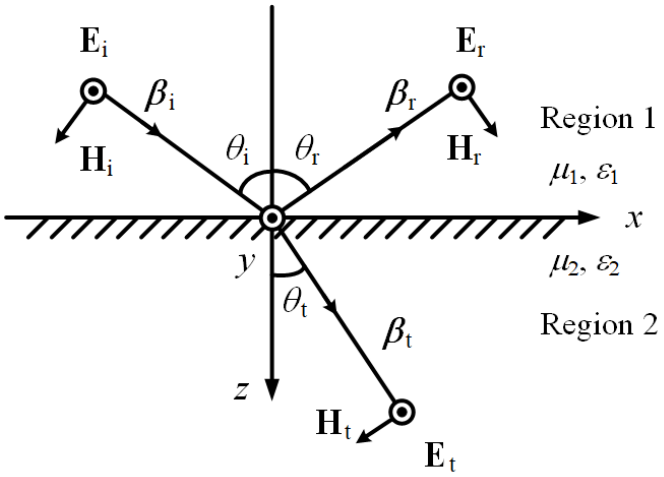


Figure 18.1: A schematic showing the reflection of the TE polarization wave impinging on a dielectric interface.

To set up the above problem, motivated by physical insight, the wave in Region 1 can be written as the superposition or sum of the incident and reflected plane waves. Here, $\mathbf{E}_i(\mathbf{r})$ and $\mathbf{E}_r(\mathbf{r})$ are the incident and reflected plane waves, respectively, with different $\boldsymbol{\beta}$ vectors. The total field is $\mathbf{E}_i(\mathbf{r})+\mathbf{E}_r(\mathbf{r})$ which are the phasor representations of the fields. We assume a plane wave polarized in the y direction where the wave vectors are $\boldsymbol{\beta}_i = \hat{x}\beta_{ix} + \hat{z}\beta_{iz}$, $\boldsymbol{\beta}_r = \hat{x}\beta_{rx} - \hat{z}\beta_{rz}$, respectively for the incident, reflected waves. In Region 2, we assume only a transmitted wave with wave vector given by $\boldsymbol{\beta}_t = \hat{x}\beta_{tx} + \hat{z}\beta_{tz}$, respectively for the transmitted waves. Then, from Section 7.3, we have, in full mathematical form, an incident plane wave is given by

$$\mathbf{E}_i = \hat{y}E_0 e^{-j\boldsymbol{\beta}_i\cdot\mathbf{r}} = \hat{y}E_0 e^{-j\beta_{ix}x - j\beta_{iz}z} \tag{18.1.1}$$

which represents a uniform incident plane wave. For the reflected plane wave, we have

$$\mathbf{E}_r = \hat{y}R^{TE}E_0 e^{-j\boldsymbol{\beta}_r\cdot\mathbf{r}} = \hat{y}R^{TE}E_0 e^{-j\beta_{rx}x + j\beta_{rz}z} \tag{18.1.2}$$

[3]These polarizations are also variously know as TE_z, or the s and p polarizations, a descendent from the notations for acoustic waves where s and p stand for shear and pressure waves, respectively.

which is a uniform reflected wave with R^{TE} being the Fresnel reflection coefficient. By the same token, the plane wave solution in Region 2 only have a transmitted plane wave; hence

$$\mathbf{E}_t = \hat{y} T^{TE} E_0 e^{-j\boldsymbol{\beta}_t \cdot \mathbf{r}} = \hat{y} T^{TE} E_0 e^{-j\beta_{tx} x - j\beta_{tz} z} \tag{18.1.3}$$

with T^{TE} being the Fresnel transmission coefficient. In the above, we assume that the incident wave is known and hence, E_0 is known. From (18.1.2) and (18.1.3), R^{TE} and T^{TE} are two unknowns yet to be sought.

To find them, we need two equations, and they can be found by imposing boundary conditions at the interface.[4] These boundary conditions are tangential $\mathbf{E}$ field continuous and tangential $\mathbf{H}$ field continuous, which become $\hat{n} \times \mathbf{E}$ continuous and $\hat{n} \times \mathbf{H}$ continuous conditions at the interface.

Imposing $\hat{n} \times \mathbf{E}$ continuous at $z = 0$, we get

$$E_0 e^{-j\beta_{ix} x} + R^{TE} E_0 e^{-j\beta_{rx} x} = T^{TE} E_0 e^{-j\beta_{tx} x}, \quad \forall x \tag{18.1.4}$$

where $\forall$ means "for all". In order for the above to be valid for all x, it is necessary that $\beta_{ix} = \beta_{rx} = \beta_{tx}$, which is also known as the phase matching condition.[5] From the above, by letting $\beta_{ix} = \beta_{rx} = \beta_1 \sin\theta_i = \beta_1 \sin\theta_r$, we obtain that $\theta_r = \theta_i$ or that the law of reflection stating that the angle of reflection is equal to the angle of incidence.

By letting $\beta_{ix} = \beta_1 \sin\theta_i = \beta_{tx} = \beta_2 \sin\theta_t$, we obtain Snell's law of refraction that $\beta_1 \sin\theta_i = \beta_2 \sin\theta_t$. (This law of refraction was also known in the Islamic world in the 900 AD to Ibn Sahl [137]).

The exponential terms or the phase terms on both sides of (18.1.4) are the same. Now, canceling common terms on both sides of the equation (18.1.4), the above simplifies to

$$1 + R^{TE} = T^{TE} \tag{18.1.5}$$

Next, to impose $\hat{n} \times \mathbf{H}$ continuous at the interface, one needs to find the $\mathbf{H}$ field using $\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H}$. Since these are plane waves, we can replace ∇ with $-j\boldsymbol{\beta}$. Then we have $\mathbf{H} = -j\boldsymbol{\beta} \times \mathbf{E}/(-j\omega\mu) = \boldsymbol{\beta} \times \mathbf{E}/(\omega\mu)$. By so doing,[6]

$$\mathbf{H}_i = \frac{\boldsymbol{\beta}_i \times \mathbf{E}_i}{\omega\mu_1} = \frac{\boldsymbol{\beta}_i \times \hat{y}}{\omega\mu_1} E_0 e^{-j\boldsymbol{\beta}_i \cdot \mathbf{r}} = \frac{\hat{z}\beta_{ix} - \hat{x}\beta_{iz}}{\omega\mu_1} E_0 e^{-j\boldsymbol{\beta}_i \cdot \mathbf{r}} \tag{18.1.6}$$

$$\mathbf{H}_r = \frac{\boldsymbol{\beta}_r \times \mathbf{E}_r}{\omega\mu_1} = \frac{\boldsymbol{\beta}_r \times \hat{y}}{\omega\mu_1} R^{TE} E_0 e^{-j\boldsymbol{\beta}_r \cdot \mathbf{r}} = \frac{\hat{z}\beta_{rx} + \hat{x}\beta_{rz}}{\omega\mu_1} R^{TE} E_0 e^{-j\boldsymbol{\beta}_r \cdot \mathbf{r}} \tag{18.1.7}$$

$$\mathbf{H}_t = \frac{\boldsymbol{\beta}_t \times \mathbf{E}_t}{\omega\mu_2} = \frac{\boldsymbol{\beta}_t \times \hat{y}}{\omega\mu_2} T^{TE} E_0 e^{-j\boldsymbol{\beta}_t \cdot \mathbf{r}} = \frac{\hat{z}\beta_{tx} - \hat{x}\beta_{tz}}{\omega\mu_2} T^{TE} E_0 e^{-j\boldsymbol{\beta}_t \cdot \mathbf{r}} \tag{18.1.8}$$

Imposing $\hat{n} \times \mathbf{H}$ continuous or H_x continuous at $z = 0$, we have

$$-\frac{\beta_{iz}}{\omega\mu_1} E_0 e^{-j\beta_{ix} x} + \frac{\beta_{rz}}{\omega\mu_1} R^{TE} E_0 e^{-j\beta_{rx} x} = -\frac{\beta_{tz}}{\omega\mu_2} T^{TE} E_0 e^{-j\beta_{tx} x} \tag{18.1.9}$$

[4]Here, we will treat this problem as a boundary value problem where the unknowns are sought from equations obtained from boundary conditions.

[5]The phase-matching condition can also be proved by taking the Fourier transform of the equation with respect to x. Among the physics community, this is also known as momentum matching, as the wavenumber of a wave is related to the momentum of the particle. Remember that the momentum of a particle is proportional to $\hbar k$.

[6]Compared to transmission line theory, we note here that field theory is a lot more complicated that will drive you daffy. That is the reason for the triumph of transmission line theory as well. Thank goodness that we have transmission line theory and circuit theory which can capture the major physics of electromagnetic theory!

As mentioned before, the phase-matching condition requires that $\beta_{ix} = \beta_{rx} = \beta_{tx}$. The dispersion relation for plane waves requires that in their respective media,

$$\beta_{ix}^2 + \beta_{iz}^2 = \beta_{rx}^2 + \beta_{rz}^2 = \omega^2\mu_1\varepsilon_1 = \beta_1^2, \qquad \text{Region 1} \tag{18.1.10}$$

$$\beta_{tx}^2 + \beta_{tz}^2 = \omega^2\mu_2\varepsilon_2 = \beta_2^2, \qquad \text{Region 2} \tag{18.1.11}$$

Since

$$\beta_{ix} = \beta_{rx} = \beta_{tx} = \beta_x \tag{18.1.12}$$

the above implies that, since β_{rx} and β_{tx} are in the same region,

$$\beta_{iz} = \beta_{rz} = \beta_{1z} \tag{18.1.13}$$

Moreover, since $\beta_1 \neq \beta_2$, $\beta_{tz} = \beta_{2z} \neq \beta_{1z}$ usually.

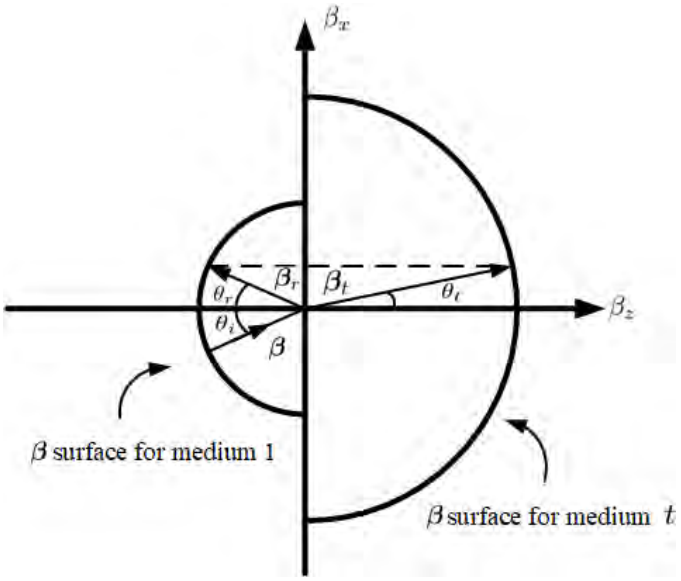


Figure 18.2: Wave-number (β or k) surfaces in two regions showing the phase matching condition. In this case, the wave number in medium t is larger (with larger $\boldsymbol{\beta}$ vector) than the wave number in medium 1. Then the wave vectors for the incident wave, reflected wave, and transmitted wave have to be aligned in such a way that their components parallel to the interface are equal because of the phase-matching condition. Also, one can see that Snell's law is satisfied as required by the phase-matching condition.

The above can be easily understood from the phase-matching diagram shown in Figure 18.2. In this figure, depending on the region that the $\boldsymbol{\beta}$ is in, the length of the $\boldsymbol{\beta}$ vector, which is $\sqrt{\beta_x^2 + \beta_z^2} = \beta = \omega\sqrt{\mu\varepsilon}$ changes from medium to medium. The circles in Figure 18.2 decide on

the length of the $\boldsymbol{\beta}$ vectors. As shown in Figure 18.2, because of the dispersion relation that $\beta_{rx}^2 + \beta_{rz}^2 = \beta_{ix}^2 + \beta_{iz}^2 = \beta_1^2$, $\beta_{tx}^2 + \beta_{tz}^2 = \beta_2^2$, they are equations of two circles in 2D whose radii are β_1 and β_2, respectively. (The tips of the $\boldsymbol{\beta}$ vectors for Regions 1 and 2 have to be on a spherical surface (also called β or k surface) in the β_x, β_y, and β_z space in the general 3D case, but in this figure, we only show a cross section of the sphere assuming that $\beta_y = 0$.

Phase matching implies that the x-component of these $\boldsymbol{\beta}$ vectors are equal to each other as shown. One sees immediately that $\theta_i = \theta_r$ in Figure 18.2, and also as θ_i increases, θ_t increases. Then (18.1.9) simplifies to

$$\frac{\beta_{1z}}{\mu_1}\left(1 - R^{TE}\right) = \frac{\beta_{2z}}{\mu_2} T^{TE} \tag{18.1.14}$$

where $\beta_{1z} = \sqrt{\beta_1^2 - \beta_x^2}$, and $\beta_{2z} = \sqrt{\beta_2^2 - \beta_x^2}$.

Solving (18.1.5) and (18.1.14) for R^{TE} and T^{TE} yields the Fresnel coefficients to be[7]

$$R^{TE} = \left(\frac{\beta_{1z}}{\mu_1} - \frac{\beta_{2z}}{\mu_2}\right) \Bigg/ \left(\frac{\beta_{1z}}{\mu_1} + \frac{\beta_{2z}}{\mu_2}\right) \tag{18.1.15}$$

$$T^{TE} = 2\left(\frac{\beta_{1z}}{\mu_1}\right) \Bigg/ \left(\frac{\beta_{1z}}{\mu_1} + \frac{\beta_{2z}}{\mu_2}\right) \tag{18.1.16}$$

[7]For mnemonics, we can also call these coefficients R_{12}^{TE} and T_{12}^{TE} with the understanding that the incident wave is in Region 1 impinging in the Regions 1 and 2 interface.

18.1.2 TM Polarization (Parallel or H Polarization)[8]

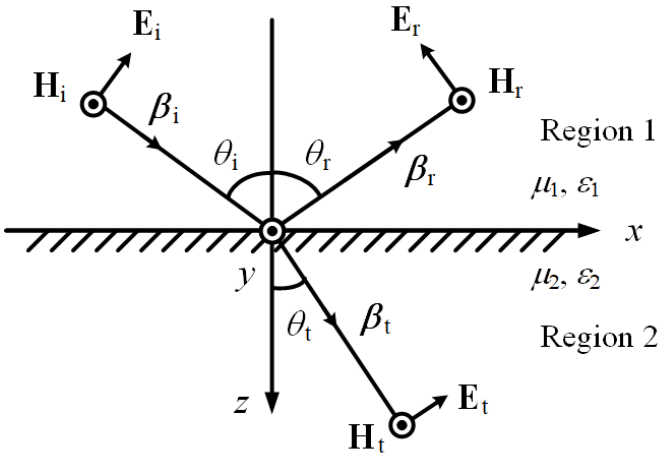


Figure 18.3: A similar schematic showing the reflection of the TM polarization wave impinging on a dielectric interface. The solution to this problem can be easily obtained from the solution for TE polarization by invoking duality principle.

The solution to the TM polarization case can be obtained by invoking the duality principle where we do the substitution $\mathbf{E} \to \mathbf{H}$, $\mathbf{H} \to -\mathbf{E}$, and $\mu \rightleftharpoons \varepsilon$ as shown in Figure 18.3. The reflection coefficient for the TM magnetic field is then

$$R^{TM} = \left(\frac{\beta_{1z}}{\varepsilon_1} - \frac{\beta_{2z}}{\varepsilon_2}\right) \Big/ \left(\frac{\beta_{1z}}{\varepsilon_1} + \frac{\beta_{2z}}{\varepsilon_2}\right) \tag{18.1.17}$$

$$T^{TM} = 2\left(\frac{\beta_{1z}}{\varepsilon_1}\right) \Big/ \left(\frac{\beta_{1z}}{\varepsilon_1} + \frac{\beta_{2z}}{\varepsilon_2}\right) \tag{18.1.18}$$

Please remember now that R^{TM} and T^{TM} are reflection and transmission coefficients for the magnetic fields, whereas R^{TE} and T^{TE} are those for the electric fields. Some textbooks may define these reflection coefficients based on electric field only, and they will look different, and duality principle cannot be applied.

18.1.3 Lens Optics and Ray Tracing

The Fresnel coefficients are derived for infinitely flat surface. But when the wavelength is very short, a curved surface resembles a flat surface to the plane wave,[9] and the Fresnel coefficients can

[8] Also known as TM_z polarization.

[9] This is very much akin to the notion that the Earth is flat to people who do not venture outside the vicinity of their neighborhood.

be used to estimate the reflected and transmitted waves. This is the fundamental principle behind lens optics.[10]

When the electromagnetic wave is described by ray, the field is known also as ray optics. In this case, Maxwell's equations are not solved in their full glory, but approximately. The approximation is a very good one when the frequency is high and the wavelength is short. We will learn more in high-frequency methods later in the course. When the geometry is simple, solving the ray optics problem is similar to solving a geometry problem since one needs to know how the rays intersect with the geometry (see Figure 18.4). But when the geometry is highly complex, ray-tracing methods are used for tracking the light rays as they propagates through a complex environment. Computer codes have been written to do ray tracing. Ray tracing has been used to enhance the fidelity of a graphical picture in the movie industry. This has in turn fueled the growth of the graphics processing units (GPU) in the computer industry [138].

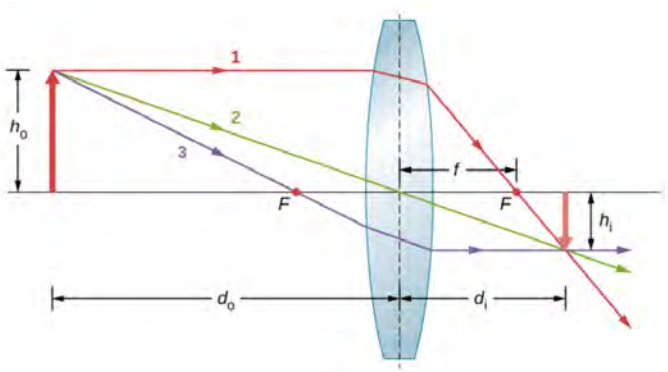


Figure 18.4: Ray tracing, based on Snell's law of refraction, can be used to solve many optics problems when the wavelength is small compared the to the size of the geometry (courtesy of Steven Mellema, The Cosmic Universe).

18.2 Interesting Physical Phenomena

Three interesting physical phenomena emerge from the solutions of the single-interface problem. They are *total internal reflection*, *Brewster angle effect*, and *surface plasmonic resonance*. We will examine them next.

[10]Lens were made in China since the Han dynasty in ancient times [136].

18.2.1 Total Internal Reflection

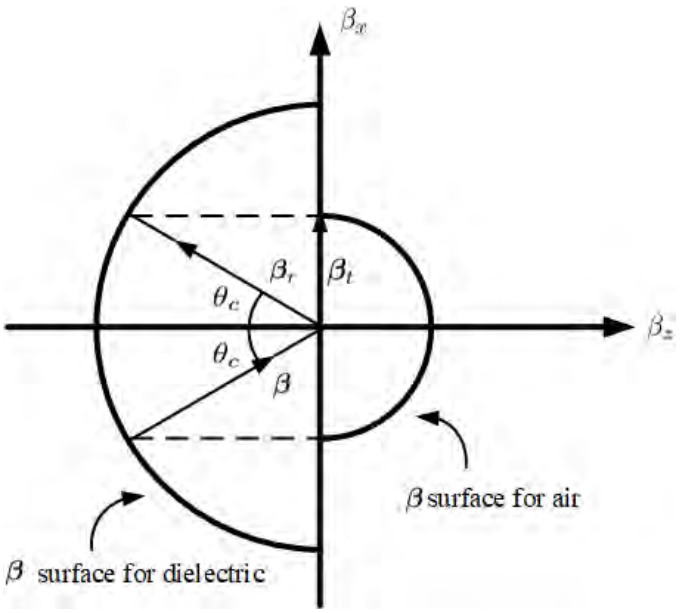


Figure 18.5: Phase-matching diagram for the total internal reflection case. Wave-number surfaces in two regions show the phase matching condition. In this case, the wave number in medium t is smaller than the wave number in medium 1. The figure shows an incident wave vector coming in at the critical angle. Then the transmitted wave vector is parallel to the interface as shown with $\beta_z = 0$. When the incident angle is larger than the critical angle, β_z becomes an imaginary number as the wave vector in Region t is complex and cannot be drawn.

How total internal reflection can come about is also explainable by phase matching via the phase-matching diagram. This phase-matching condition can be illustrated using β-surfaces as shown in Figure 18.5. It turns out that because of the phase matching condition, for interfaces where the left-hand side is optically more dense than the right-hand side, and for certain value of the incident angle, β_{2z} becomes pure imaginary.

For an optically less dense medium for which $\beta_2 < \beta_1$, according to the Snell's law of refraction, the transmitted $\boldsymbol{\beta}$ will refract away from the normal, as seen in the figure. Therefore, eventually the vector $\boldsymbol{\beta}_t$ becomes parallel to the x axis when $\beta_{ix} = \beta_{rx} = \beta_2 = \omega\sqrt{\mu_2\varepsilon_2}$ and $\theta_t = \pi/2$. The incident angle at which this happens is termed the critical angle θ_c (see Figure 18.5).

At the critical angle, since $\beta_{ix} = \beta_1 \sin\theta_i = \beta_{rx} = \beta_1 \sin\theta_r = \beta_2$, or

$$\sin\theta_r = \sin\theta_i = \sin\theta_c = \frac{\beta_2}{\beta_1} = \frac{\sqrt{\mu_2\varepsilon_2}}{\sqrt{\mu_1\varepsilon_1}} = \frac{n_2}{n_1} \tag{18.2.1}$$

where n_1 is the refractive index defined as $c_0/v_i = \sqrt{\mu_i \varepsilon_i}/\sqrt{\mu_0 \varepsilon_0}$ where v_i is the phase velocity of the wave in Region i. Hence,

$$\theta_c = \sin^{-1}(n_2/n_1) \tag{18.2.2}$$

When $\theta_i > \theta_c$. $\beta_x > \beta_2$ and $\beta_{2z} = \sqrt{{\beta_2}^2 - {\beta_x}^2}$ becomes pure imaginary. When β_{2z} becomes pure imaginary, the wave cannot propagate in Region 2, or $\beta_{2z} = -j\alpha_{2z}$, and the wave becomes evanescent. The physical reason for the decaying nature of the evanescent wave is quite different from that of a decaying wave in a lossy medium. The former is due to phase matching, and the need for the field to satisfy the boundary condition, while the latter is due to the loss of energy to the lossy medium. In a lossless medium, one can also show that the evanescent wave does not carry real power, but only reactive power using the complex Poynting's theorem.

The reflection coefficient (18.1.15) becomes of the form

$$R^{TE} = (A - jB)/(A + jB) \tag{18.2.3}$$

Since the numerator is the complex conjugate of the denominator. It is clear that $|R^{TE}| = 1$ always, and that $R^{TE} = e^{j\theta_{TE}}$. Therefore, a total internally reflected wave suffers a phase shift. A phase shift in the frequency domain corresponds to a time delay in the time domain. Such a time delay is achieved by the wave traveling laterally in Region 2 before being refracted back to Region 1. Such a lateral shift is called the Goos-Hanschen shift as shown in Figure 18.6 [61].[11] A wave that travels laterally along the surface of two media is also known as lateral waves [139, 140].

(Please be reminded that total internal reflection comes about entirely due to the phase-matching condition when Region 2 is a faster medium than Region 1. Hence, it will occur with all manner of waves, such as elastic waves, sound waves, seismic waves, quantum waves etc., and even waves in a cylindrical fiber.)

[11] You may be perplexed by our use of finite beam width of the plane wave for our physical argument. But you will learn later that a finite beam width can be approximated by a bundle of plane waves with similar wave numbers.

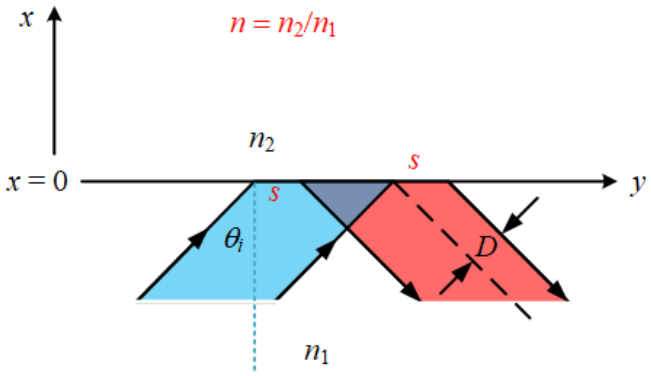


Figure 18.6: Goos-Hanschen Shift. A phase delay is equivalent to a time delay. In the time domain, one imagines a pulse traveling along a ray. When the ray hits the interface, it travels laterally on the interface for a distance s before being refracted back. The extra distance s travelled gives rise to the time delay (courtesy of Paul R. Berman (2012), Scholarpedia, 7(3):11584 [141]).

The guidance of a wave in a dielectric slab is due to total internal reflection at the dielectric-to-air interface. The wave bounces between the two interfaces of the slab, and creates evanescent waves outside, as shown in Figure 18.7. The guidance of waves in an optical fiber works by a similar mechanism of total internal reflection, as shown in Figure 18.8. Optical light wave can provide carrier frequencies that are extremely high compared to radio or microwave, giving rise to the tremendous bandwidth increase. Due to the tremendous impact the optical fiber has on modern-day communications, Charles Kao, the father of the optical fiber, was awarded the Nobel Prize in 2009. His work was first published in [142].

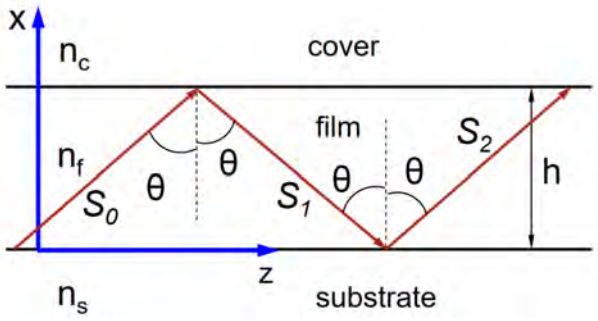


Figure 18.7: The total internal reflections (TIR) at the two interfaces of a thin-film waveguide can be used to guide an optical wave (courtesy of E.N. Glytsis, NTUA, Greece [143]).

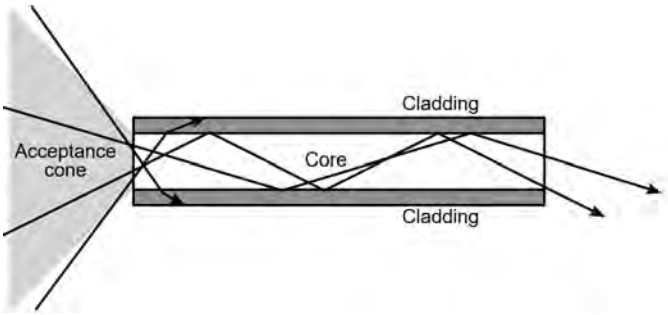


Figure 18.8: An optical fiber consists of a core and a cladding region. Total internal reflections (TIR), due to phase matching condition, can occur at the core-cladding interface. TIR can guide an optical wave in the fiber which has a tremendous impact on the communication industry due to the low loss of the optical fiber (courtesy of Wikepedia [144]).

Waveguides have affected international communications for over a hundred year now. Since telegraphy was in place before the full advent of Maxwell's equations, submarine cables for global communications were laid as early as 1850's. Figure 18.9 shows a submarine cable from 1869 using coaxial cable, compared to the one used in the modern world using optical fiber!

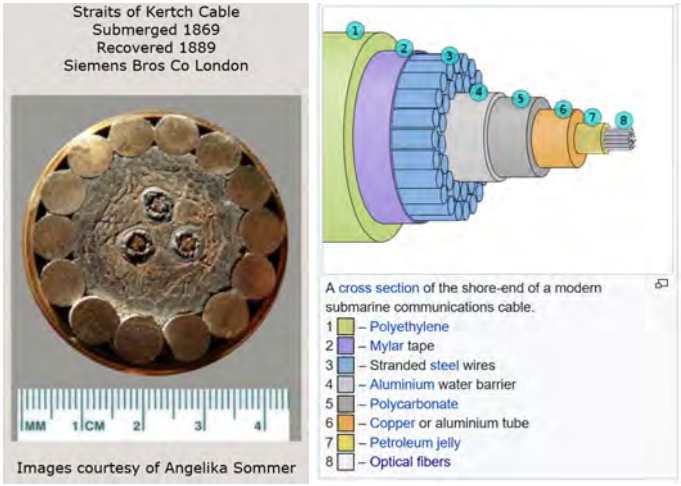


Figure 18.9: The picture of an old 1869 submarine cable made of coaxial cables (left), and modern submarine cable made of optical fibers (right) (courtesy of Atlantic-Cable [145], and Wikipedia [146]). Comparing them to the optical fiber now, we have come a long way!

Exercises for Lecture 18

Problem 18-1:

(i) Derive the Fresnel reflection coefficients for TE and TM plane waves for the single interface problem.

(ii) Explain how Snell's law follows from the phase matching condition.

(iii) Find the condition for total internal reflection, and explain why the wave is evanescent on one side of the interface. Under what condition can total internal reflection occur, and in which medium is the wave evanescent?

(iv) For the lossless medium case, what happens to the amplitude of the reflection coefficient when total internal reflection occurs?

(v) Using the complex Poynting's theorem, show that the evanescent wave can only carry reactive power in a lossless medium.

Chapter 19

Brewster Angle, SPP, Homomorphism with Transmission Lines

Though simple that it looks, embedded in the TM Fresnel reflection coefficient are a few more interesting physical phenomena. We have looked at the physics of total internal reflection, which has inspired many interesting technologies such as waveguides, the most important of which is the optical fiber. In this lecture, we will look at other physical phenomena. These are the phenomena of Brewster's angle [147, 148] and that of surface plasmon resonance, or polariton [149, 150].

Even though transmission line theory and the theory of plane wave reflection and transmission look quite different, they are very similar in their underlying mathematical structures. For lack of a better name, we call this mathematical "homomorphism" (math analogy).[1] Later, to simplify the mathematics of waves in layered media, we will draw upon this mathematical "homomorphism" between multi-section transmission line theory and plane-wave theory in layered media.

19.1 Brewster's Angle

First, we will continue with understanding some interesting phenomena associated with the single-interface problem starting with the Brewster's angle. Brewster angle was discovered in 1815 [147, 148]. Furthermore, most materials at optical frequencies have $\varepsilon_2 \neq \varepsilon_1$, but $\mu_2 \approx \mu_1$. In other words, it is hard to obtain magnetic materials at optical frequencies. Therefore, even though the TE and TM polarizations are dual to each other, the TM polarization for light behaves differently from TE polarization. As such, we shall focus on the reflection and transmission of the TM polarization

[1] The use of this term could be to the chagrin of a math person, but it has also being used in a subject called homomorphic encryption or computing [151].

of light: The previously derived TM reflection coefficient are given here, viz.,

$$R^{TM} = \left(\frac{\beta_{1z}}{\varepsilon_1} - \frac{\beta_{2z}}{\varepsilon_2}\right) \Big/ \left(\frac{\beta_{1z}}{\varepsilon_1} + \frac{\beta_{2z}}{\varepsilon_2}\right) \tag{19.1.1}$$

The transmission coefficient is easily gotten by the formula $T^{TM} = 1 + R^{TM}$.

Observe that for R^{TM}, it is possible that $R^{TM} = 0$ if

$$\varepsilon_2 \beta_{1z} = \varepsilon_1 \beta_{2z} \tag{19.1.2}$$

Squaring the above, making the note that $\beta_{iz} = \sqrt{\beta_i^2 - \beta_x^2}$, one gets[2]

$$\varepsilon_2{}^2(\beta_1{}^2 - \beta_x{}^2) = \varepsilon_1{}^2(\beta_2{}^2 - \beta_x{}^2) \tag{19.1.3}$$

where β_x is the same for all regions because of phase matching. Solving the above, assuming $\mu_1 = \mu_2 = \mu$, gives

$$\beta_x = \omega\sqrt{\mu}\sqrt{\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1 + \varepsilon_2}} = \underbrace{\beta_1 \sin\theta_1}_{\beta_{1x}} = \underbrace{\beta_2 \sin\theta_2}_{\beta_{2x}} \tag{19.1.4}$$

The latter two equalities in the above come from phase matching. Therefore, at the Brewster angle, by letting $\beta_i = \omega\sqrt{\mu\varepsilon_i}$ in the above, we get

$$\sin\theta_1 = \sqrt{\frac{\varepsilon_2}{\varepsilon_1 + \varepsilon_2}}, \quad \sin\theta_2 = \sqrt{\frac{\varepsilon_1}{\varepsilon_1 + \varepsilon_2}} \tag{19.1.5}$$

or squaring the above and adding them, we get

$$\sin^2\theta_1 + \sin^2\theta_2 = 1, \tag{19.1.6}$$

Then, assuming that θ_1 and θ_2 are less than $\pi/2$, and using the identity that $\cos^2\theta_1 + \sin^2\theta_1 = 1$, after subtracting this identity from the above, we infer that $\cos^2\theta_1 = \sin^2\theta_2$. It follows that

$$\sin\theta_2 = \cos\theta_1 \tag{19.1.7}$$

In other words, because $\sin\theta_2 = \cos(\frac{\pi}{2} - \theta_2)$, the above implies that

$$\theta_1 + \theta_2 = \pi/2 \tag{19.1.8}$$

The above formula can be used to explain why at Brewster angle, no light is reflected back to Region 1. Figure 19.1 shows that the induced polarization dipoles in Region 2 always have their axes aligned in the direction of reflected wave. A dipole does not radiate along its axis, a fact that we will learn later in the course. But at this point, it can be verified heuristically by field sketch and looking at the Poynting's vector. Therefore, these induced dipoles in Region 2 do not radiate in the direction of the reflected wave. (Notice that when the contrast is very weak meaning that $\varepsilon_1 \cong \varepsilon_2$, then $\theta_1 \cong \theta_2$. And from (19.1.8), we infer that $\theta_1 \cong \theta_2 \cong \pi/4$.)

[2]Notice that if $\varepsilon_2 = \varepsilon_1$, which is possible for the TE polarization where these will be for permeability for non-magnetic materials, then the solution is $\beta_x = \infty$ which is an uninteresting solution.

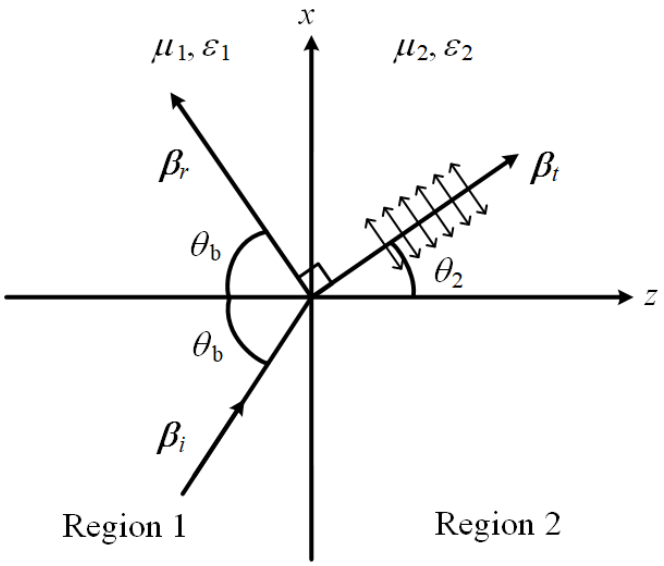


Figure 19.1: A figure showing a plane wave being reflected and transmitted at the Brewster's angle. In Region t (or Region 2), the transmitted field induces small dipoles to produce polarization density $\mathbf{P}$ with polarization current $j\omega\mathbf{P}$. The polarization current or dipoles are all pointing in the β_r direction, and hence, there is no radiation in that direction. In this figure, $\theta_b = \theta_1$, and hence, from (19.1.8), $\theta_b + \theta_2 = \pi/2$.

Because of the Brewster angle effect for TM polarization when $\varepsilon_2 \neq \varepsilon_1$, $|R^{TM}|$ has to go through a null when $\theta_i = \theta_b$. Therefore in general, $|R^{TM}| \leq |R^{TE}|$ is as shown in the plots in Figure 19.2. Thus when a randomly (or arbitrarily) polarized light is incident on a surface, the polarization where the electric field is parallel to the surface (TE polarization) is reflected more than the polarization where the magnetic field is parallel to the surface (TM polarization). This phenomenon is used to design sun glasses to reduce road surface glare for drivers. For light reflected off a road surface, they are predominantly horizontally polarized with respect to the surface of the road. When sun glasses are made with vertical polarizers,[3] they will filter out or mitigate the reflected rays from the road surface to reduce road glare. This phenomenon can also be used to improve the quality of photography by using a polarizer filter as shown in Figure 19.3.

[3]Defined as one that will allow vertical polarization to pass through.

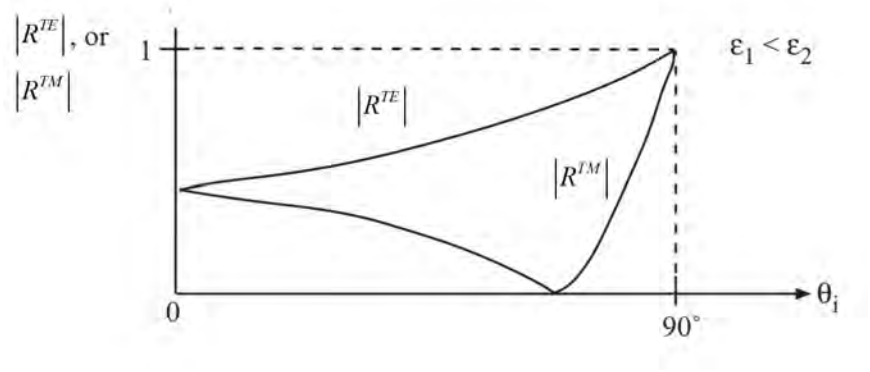


Figure 19.2: Because $|R^{TM}|$ has to go through a null when $\theta_i = \theta_b$; therefore, $|R^{TM}| \leq |R^{TE}|$ for all θ_i as shown in the above plots.

Figure 19.3: Because that TM and TE lights will be reflected differently, polarizer filter can produce remarkable effects on the quality of the photograph by reducing glare [148].

19.1.1 Surface Plasmon Polariton (SPP)

Surface plasmon polariton (SPP) occurs for the same mathematical reason for the Brewster angle effect but the physical mechanism is quite different. Many papers and textbooks will introduce this phenomenon from a different angle. But here, we will see it from the Fresnel reflection coefficient for the TM waves. When the denominator of the reflection coefficient R^{TM} is zero, it can become infinite. (This heralds the presence of some interesting physical phenomena, and in this case, a resonance behavior.) This is possible if $\varepsilon_2 < 0$, which is possible if medium 2 is a plasma medium. In this case, the criterion for the denominator to be zero is

$$-\varepsilon_2 \beta_{1z} = \varepsilon_1 \beta_{2z} \tag{19.1.9}$$

When R^{TM} becomes infinite, it implies that a reflected wave exists with no incident wave.[4] Hence, there is a plasmonic resonance or guided mode existing at the interface without the presence of an incident wave. It is a self-sustaining wave propagating in the x direction; and hence, it is a guided mode propagating in the x direction.

Solving (19.1.9) after squaring it, as in the Brewster angle case, yields

$$\beta_x = \omega\sqrt{\mu}\sqrt{\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1+\varepsilon_2}} \tag{19.1.10}$$

This is exactly the same equation for the Brewster angle except now that ε_2 is negative.[5] Even though $\varepsilon_2 < 0$, but $\varepsilon_1 + \varepsilon_2 < 0$ is still possible, so that the expression under the square root sign in (19.1.10) is positive. Thus, β_x can be pure real. The corresponding β_{1z} and β_{2z} in (19.1.9) can be pure imaginary as explained below, and (19.1.9) can still be satisfied.

We shall show that the above corresponds to a guided wave propagating in the x direction. When this happens,

$$\beta_{1z} = \sqrt{{\beta_1}^2 - {\beta_x}^2} = \omega\sqrt{\mu}\left[\varepsilon_1\left(1 - \frac{\varepsilon_2}{\varepsilon_1+\varepsilon_2}\right)\right]^{1/2} = \omega\sqrt{\mu}\left[\frac{\varepsilon_1^2}{\varepsilon_1+\varepsilon_2}\right]^{1/2} \tag{19.1.11}$$

Since $\varepsilon_1 + \varepsilon_2 < 0$, it is seen that β_{1z} becomes pure imaginary. Moreover, $\beta_{2z} = \sqrt{{\beta_2}^2 - {\beta_x}^2}$ and ${\beta_2}^2 < 0$ making β_{2z} becomes even a larger imaginary number. This corresponds to a trapped wave (or a bound state) at the interface. The wave decays exponentially in both directions away from the interface and they are both evanescent waves.[6] The physical characteristic of this mode is shown in Figure 19.4, and is the only case in electromagnetics where a single interface can guide a surface wave, while such phenomenon abounds for elastic waves [152, 153].

When one operates close to the resonance of the mode so that the denominator in (19.1.10) is almost zero, then β_x can be very large. The wavelength in the x direction becomes very short in this case. And since $\beta_{iz} = \sqrt{\beta_i^2 - \beta_x^2}$, then β_{1z} and β_{2z} become even larger imaginary numbers. Thus the mode becomes tightly confined or bound to the interface, making the confinement of the mode very tight. This evanescent wave is much more rapidly decaying than that encountered in total internal reflection, at an interface between medium 1 and medium 2. There, $\beta_{2z} = \sqrt{\beta_2^2 - \beta_x^2}$ where β_x is no larger than β_1. On the other hand, the SPP mode is tightly confined to the interface. It portends its use in tightly packed optical components where cross-talk between components can be an issue. As such, this has stirred up some excitement in the optics community.

An ordinary waveguide like a transmission line requires the exchange between stored electric energy and magnetic energy for the guided wave. But for this SPP waveguide, the magnetic field is conspicuously absent: it is the kinetic energy of the electrons in the plasma that provides the other stored energy component; and hence, the waveguide can be made small. The resonance is

[4] In other words, a solution exists without the excitation term. This is often encountered in a resonance system like an LC tank circuit. Current flows in the tank circuit despite the absence of an exciting or driving voltage. In an ordinary differential equation or partial differential equation without a driving term (source term), such solutions are known as homogeneous solutions (to clarify the potpouri of math terms, homogeneous solutions here refer to a solution with zero source term). In a matrix equation $\overline{\mathbf{A}} \cdot \mathbf{x} = \mathbf{b}$ without a right-hand side or that $\mathbf{b} = 0$, it is known as a null-space solution. These are all equivalent ways of saying a resonance solution.

[5] We see that it is often a dangerous proposition to square and square-root a function. We have to be guided by physical insight to see if we are finding a sane or insane solution!

[6] We have learnt about evanescent waves in Lecture 18. These waves do not carry real power.

due to the exchange of the stored electric field energy and the kinetic energy of the electrons. Since the kinetic energy of the electrons replaces the stored energy of the magnetic field in an ordinary waveguide, this effect is also known as the kinetic inductance.

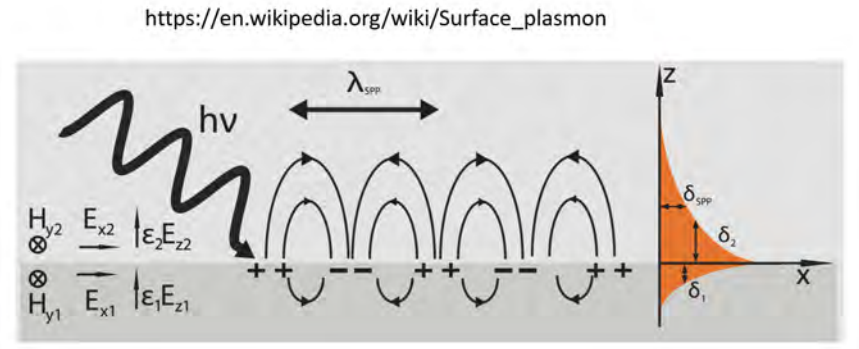


Figure 19.4: Figure showing a surface plasmonic mode propagating at an air-plasma interface. As in all resonant systems, a resonant mode entails the exchange of energies. In the case of surface plasmonic resonance, the energy is exchanged between the kinetic energy of the electrons (also known as kinetic inductance) and the energy store in the electric field (courtesy of Wikipedia [154]).

19.2 "Homomorphism" of Uniform Plane Waves and Transmission Lines Equations

Transmission line theory is very simple due to its one-dimensional nature. But the problem of reflection and transmission of plane waves at a planar interface is actually "homormophic" to that of the transmission line problem. Therefore, the problem of plane waves through layered medium can be mapped into the multi-section transmission line problem due to this mathematical "homomorphism" between the two problems. Hence, we can kill two birds with one stone: apply all the trappings of transmission line theory that we have learnt to solve for the solutions of waves through layered medium problems.[7] Transmission line theory is simple and well loved by engineers.

For uniform plane waves, since they are proportional to $\exp(-j\boldsymbol{\beta}\cdot\mathbf{r})$, we know that with $\nabla \to -j\boldsymbol{\beta}$, Maxwell's equations for a general isotropic homogeneous medium become

$$\boldsymbol{\beta}\times\mathbf{E} = \omega\mu\mathbf{H} \tag{19.2.1}$$

$$\boldsymbol{\beta}\times\mathbf{H} = -\omega\varepsilon\mathbf{E} \tag{19.2.2}$$

We will specialize these equations for different polarizations.

[7] This treatment is not found elsewhere, and is peculiar to these lecture notes.

19.2.1 TE or TE$_z$ Waves

For this, one assumes a TE wave traveling in the z direction with electric field polarized in the y direction, or $\mathbf{E} = \hat{y}E_y$. The corresponding magnetic field can be derive, and is seen in Figure 18.1, $\mathbf{H} = \hat{x}H_x + \hat{z}H_z$. Then we have from (19.2.1) that

$$\beta_z E_y = -\omega\mu H_x \tag{19.2.3}$$

$$\beta_x E_y = \omega\mu H_z \tag{19.2.4}$$

From (19.2.2), we deduce that

$$\beta_z H_x - \beta_x H_z = -\omega\varepsilon E_y \tag{19.2.5}$$

The above equations involve three variables, E_y, H_x, and H_z. Our goal is to make these equations homomorphic to the telegrapher's equations. But there are only two variables in the telegrapher's equations which are V and I. To this end, we will eliminate one of the variables from the above three equations. Using (19.2.4), we can express H_z in terms of E_y. Then, we can show from (19.2.5) that

$$\begin{aligned}\beta_z H_x &= -\omega\varepsilon E_y + \beta_x H_z = -\omega\varepsilon E_y + \frac{\beta_x^2}{\omega\mu} E_y \\ &= -\omega\varepsilon(1 - \beta_x^2/\beta^2)E_y = -\omega\varepsilon\cos^2\theta E_y\end{aligned} \tag{19.2.6}$$

where $\beta_x = \beta\sin\theta$ has been used.

Since we still have a plane wave, Eqns. (19.2.3) and (19.2.6) can be written to look like the telegrapher's equations by letting $-j\beta_z \to d/dz$. Thus, we have

$$\frac{d}{dz}E_y = j\omega\mu H_x \tag{19.2.7}$$

$$\frac{d}{dz}H_x = j\omega\varepsilon\cos^2\theta E_y \tag{19.2.8}$$

If we let $E_y \to V$, $H_x \to -I$, $\mu \to L$, $\varepsilon\cos^2\theta \to C$, the above is exactly analogous to the telegrapher's equations. The equivalent characteristic impedance of these equations above is then

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{\mu}{\varepsilon}}\frac{1}{\cos\theta} = \sqrt{\frac{\mu}{\varepsilon}}\frac{\beta}{\beta_z} = \frac{\omega\mu}{\beta_z} \tag{19.2.9}$$

The above $\omega\mu/\beta_z$ is also known as the wave impedance for a propagating plane wave with propagation direction or the $\boldsymbol{\beta}$ vector inclined with an angle θ respect to the z axis. It is analogous to the characteristic impedance Z_0 of a transmission line. When $\theta = 0$, as can be shown, the wave impedance $\omega\mu/\beta_z$ becomes the intrinsic impedance of space.

A two-region, single-interface reflection problem can then be mathematically mapped to a single-junction connecting two-transmission-lines problem discussed in Section 17.1.1. The equivalent characteristic impedances of these two regions are then

$$Z_{01} = \frac{\omega\mu_1}{\beta_{1z}}, \quad Z_{02} = \frac{\omega\mu_2}{\beta_{2z}} \tag{19.2.10}$$

We can use the above to find Γ_{12} as given by

$$\Gamma_{12} = \frac{Z_{02} - Z_{01}}{Z_{02} + Z_{01}} = \frac{(\mu_2/\beta_{2z}) - (\mu_1/\beta_{1z})}{(\mu_2/\beta_{2z}) + (\mu_1/\beta_{1z})} \tag{19.2.11}$$

The above is the same as the Fresnel reflection coefficient we have previously derived for TE waves or R^{TE} after some simple re-arrangement, but they are derived differently here.

Assuming that we have a single junction transmission line, one can define a transmission coefficient given by

$$T_{12} = 1 + \Gamma_{12} = \frac{2Z_{02}}{Z_{02} + Z_{01}} = \frac{2(\mu_2/\beta_{2z})}{(\mu_2/\beta_{2z}) + (\mu_1/\beta_{1z})} \tag{19.2.12}$$

The above is similar to the continuity of the voltage across the junction, which is the same as the continuity of the tangential electric field across the interface. It is also the same as the Fresnel transmission coefficient T^{TE}.

19.2.2 TM or TM$_z$ Waves

For TM polarization, by invoking duality principle, the corresponding equations are, from (19.2.7) and (19.2.8),

$$\frac{d}{dz}H_y = -j\omega\varepsilon E_x \tag{19.2.13}$$

$$\frac{d}{dz}E_x = -j\omega\mu\cos^2\theta H_y \tag{19.2.14}$$

Just for consistency of units, since electric field is in V m^{-1}, and magnetic field is in A m^{-1} we may chose the following map to convert the above into the telegrapher's equations, viz;

$$E_y \to V, \quad H_y \to I, \quad \mu\cos^2\theta \to L, \quad \varepsilon \to C \tag{19.2.15}$$

Then, the equivalent characteristic impedance is now

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{\mu}{\varepsilon}}\cos\theta = \sqrt{\frac{\mu}{\varepsilon}}\frac{\beta_z}{\beta} = \frac{\beta_z}{\omega\varepsilon} \tag{19.2.16}$$

The above is also termed the wave impedance of a TM propagating wave making an inclined angle θ with respect to the z axis. Notice that this wave impedance again becomes the intrinsic impedance of space when $\theta = 0$.

Now, using the reflection coefficient for a single-junction transmission line, and the appropriate characteristic impedances for the two lines as given in (19.2.16), we arrive at

$$\Gamma_{12} = \frac{(\beta_{2z}/\varepsilon_2) - (\beta_{1z}/\varepsilon_1)}{(\beta_{2z}/\varepsilon_2) + (\beta_{1z}/\varepsilon_1)} \tag{19.2.17}$$

Notice that (19.2.17) has a sign difference from the definition of R^{TM} derived earlier in the last lecture. The reason is that R^{TM} defined previously is for the reflection coefficient of magnetic field

while Γ_{12} above is for the reflection coefficient of the voltage or the electric field. This difference is also seen in the definition for transmission coefficients.[8] A voltage transmission coefficient can be defined to be

$$T_{12} = 1 + \Gamma_{12} = \frac{2(\beta_{2z}/\varepsilon_2)}{(\beta_{2z}/\varepsilon_2) + (\beta_{1z}/\varepsilon_1)} \tag{19.2.18}$$

But this will be the transmission coefficient for the voltage, which is not the same as T^{TM} which is the transmission coefficient for the magnetic field or the current. Different textbooks may define different transmission coefficients for this polarization.

[8] This is often the source of confusion for these reflection and transmission coefficients.

Exercises for Lecture 19

Problem 19-1:

(i) Explain the criterion for Brewster angle, and explain why it is prevalent for TM waves.

(ii) Explain heuristically why a electric dipole does not radiate in their axial directions, or along the axis of the dipole. Hint: Sketch the electric and magnetic fields in the vicinity of an electric dipole, and look at the Poynting's vector.

(iii) Explain why the condition for surface plasmonic polariton (SPP) is the same as the condition for Brewster angle. Then how does one distinguish an SPP case from the Brewster angle case.

(iv) Explain why the SPP mode is tightly confined to the interface, more so than the evanescent wave due to total internal reflection.

(v) Explain why the reflection and transmission from a single interface problem can be made "homomorphic" to a single-junction transmission line problem.

Chapter 20

Waves in Layered Media

"Waves in layered media" is an important topic in electromagnetics. Many media can be approximated by planarly layered media. For instance, the propagation of radio wave on the earth surface was of interest and first tackled by Sommerfeld in 1909 [155]. The earth can be approximated by planarly layered media to capture the important physics behind the wave propagation. As such, many geophysics problems can be understood by studying waves in layered media. Many microwave components are made by planarly layered structures or laminated structures such as microstrip and coplanar waveguides. Layered media are also important in optics: they can be used to make optical filters such as Fabry-Perot filters. As technologies and fabrication techniques become better, there is an increasing need to understand the interaction of waves with layered structures.

20.1 Waves in Layered Media

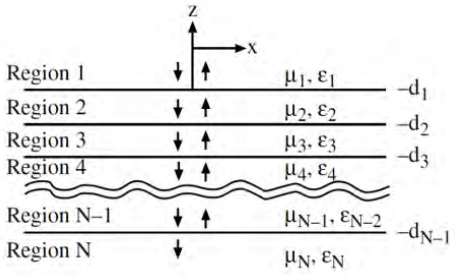


Figure 20.1: Waves in layered media. A wave entering the medium from above can be multiply reflected before emerging from the top again or transmitted to the bottom-most medium.

20.1.1 Composite Reflection Coefficient for Layered Media

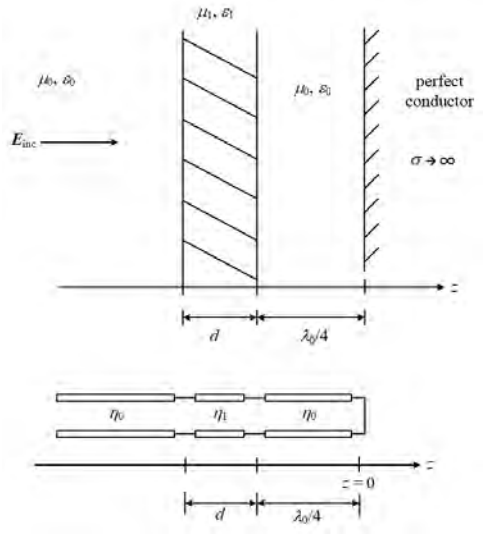


Figure 20.2: The equivalence of a layered medium problem to a transmission line problem. This equivalence is possible even for oblique incidence where we define a wave impedance as defined in the previous lecture. The wave impedance is analogous to the characteristic impedance of a transmission line. For normal incidence, the wave impedance defined in the previous lecture becomes intrinsic impedances (courtesy of J.A. Kong, Electromagnetic Wave Theory).

Because of the "homomorphism" between the transmission line problem and the plane-wave reflection by interfaces, we will exploit the simplicity of the transmission line theory to arrive at formulas for plane wave reflection by layered media. We can capitalize on using the multi-section transmission line formulas for composite reflection coefficient, which is

$$\tilde{\Gamma}_{12} = \frac{\Gamma_{12} + \tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}}{1 + \Gamma_{12}\tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}} \tag{20.1.1}$$

In the above, Γ_{12} is the local reflection at the 1,2 junction, whereas $\tilde{\Gamma}_{ij}$ are the composite reflection coefficient at the i, j interface. For instance, $\tilde{\Gamma}_{12}$ includes multiple reflections from behind the 1,2 junction. It can be used to study electromagnetic waves in layered media shown in Figures 20.1 and 20.2.

Using the result from the multi-junction transmission line, by analogy where characteristic impedances are equivalent to wave impedances, we can write down the composite reflection coefficient for a layered medium with an incident wave at the 1,2 interface, including multiple reflections from behind the interface. In addition to the wave-impedance to characteristic-impedance replace-

ments,[1] we do the following replacements: $\Gamma_{12} \to R_{12}$, $\tilde{\Gamma}_{23} \to \tilde{R}_{23}$, $\tilde{\Gamma}_{12} \to \tilde{R}_{12}$, and $\beta_2 \to \beta_{2z}$. Then we have

$$\tilde{R}_{12} = \frac{R_{12} + \tilde{R}_{23} e^{-2j\beta_{2z} l_2}}{1 + R_{12}\tilde{R}_{23} e^{-2j\beta_{2z} l_2}} \tag{20.1.2}$$

where R_{12} is the local Fresnel reflection coefficient and $\tilde{R}_{ij}$ is the composite reflection coefficient at the i, j interface. Here, l_2 is now the thickness of the region 2. In the above, we assume that the wave is incident from medium (region) 1 which is semi-infinite, the composite reflection coefficient $\tilde{R}_{12}$ above is defined at the media 1 and 2 interface. It is assumed that there are multiple reflections coming from the right of the 2,3 interface, so that the 2,3 reflection coefficient is the composite reflection coefficient $\tilde{R}_{23}$.

Figure 20.2 shows the case of a normally incident wave into a layered media. For this case, the wave impedance, defined in the previous lecture, becomes the intrinsic impedance of homogeneous space.

20.1.2 Ray Series Interpretation of Composite Reflection Coefficient

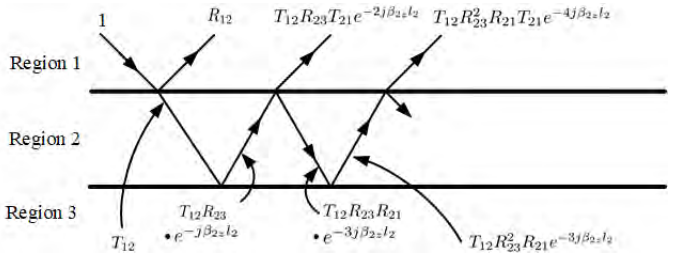


Figure 20.3: The expression of the composite reflection coefficient into a ray series. Here, $l_2 = d_2 - d_1$ is the thickness of the slab (courtesy of [108]).

For simplicity, we will assume that $\tilde{R}_{23} = R_{23}$ in this section. By manipulation, one can convert the composite reflection coefficient $\tilde{R}_{12}$ into a form that has a ray physics interpretation. By adding and subtracting the term

$$R_{12}^2 R_{23} e^{-2j\beta_{2z} l_2}$$

on the numerator of (20.1.2), and rearranging terms later, (20.1.2) can be shown to become

$$\tilde{R}_{12} = R_{12} + \frac{R_{23} e^{-2j\beta_{2z} l_2}(1 - R_{12}^2)}{1 + R_{12}R_{23} e^{-2j\beta_{2z} l_2}} \tag{20.1.3}$$

By using the fact that $R_{12} = -R_{21}$, which can be easily seen from the Fresnel reflection coefficient formula. Now we define a transmission coefficient such that $T_{ij} = 1 + R_{ij}$. Making $1 - R_{12}^2 =$

[1] Remember from the previous lecture that the wave impedance formulas are different for TM and TE polarizations.

$(1-R_{12})(1+R_{12}) = (1-R_{12})(1-R_{21}) = T_{12}T_{21}$, then the above can be rewritten as

$$\tilde{R}_{12} = R_{12} + \frac{T_{12}T_{21}R_{23}e^{-2j\beta_{2z}l_2}}{1+R_{12}R_{23}e^{-2j\beta_{2z}l_2}} \tag{20.1.4}$$

Next using the fact that $(1-x)^{-1} = 1+x+x^2+\ldots+$, the denominator of the second term above can be expanded as a power series, and the above can further be rewritten as

$$\tilde{R}_{12} = R_{12} + T_{12}R_{23}T_{21}e^{-2j\beta_{2z}l_2} + T_{12}R_{23}^2R_{21}T_{21}e^{-4j\beta_{2z}l_2} + \cdots . \tag{20.1.5}$$

In the above, the series helps us to elucidate the physics of the composite reflection coefficient $\tilde{R}_{12}$. The first term in the above is just the result of a single reflection off the first interface. The n-th term above is the consequence of the n-th reflection from the three-layer medium (see Figure 20.3). Hence, the expansion of (20.1.2) into (20.1.5) renders a lucid physical interpretation for the composite reflection coefficient. Consequently, the series in (20.1.5) can be thought of as a *ray* series or a *geometrical optics* series. It is the consequence of multiple reflections and transmissions in region 2 of the three-layer medium. It is obtained by expanding the denominator of the second term in (20.1.4). Hence, the denominator of the second term in (20.1.4) can be physically interpreted as a consequence of multiple reflections within region 2.

20.2 Phase Velocity and Group Velocity

We have seen that a single interface can guide a wave as in the case of the surface plasmon polariton (SPP, see Section 19.1.1). Later, you will see that multiple interfaces are used to guide waves as in the dielectric slab waveguide. These guided waves will have a phase velocity that we have seen before. But if a multi-frequency pulse is send along these waveguides, the pulse will not travel at the phase velocity. Instead, it travels at the group velocity as we shall see.

In general, a medium can be frequency dispersive in a complicated fashion as in the Drude-Lorentz-Sommerfeld (DLS) model. A monochromatic signal will travel with a phase velocity as we shall see. On the other hand, a polychromatic signal will propagate with a group velocity. We are ready to investigate the difference between the phase velocity and the group velocity. In this course, we will use k and β interchangeably to represent wavenumber.[2]

20.2.1 Phase Velocity

The phase velocity is the velocity of the phase of a wave. It is only defined for a monochromatic signal (also called time-harmonic, CW—constant wave, or sinusoidal signal) at one given frequency. Given a sinusoidal wave signal, e.g., the voltage signal on a transmission line, using phasor technique, its representation in the time domain can be easily found and take the form

$$\begin{aligned} V(z,t) &= V_0\cos(\omega t - kz + \alpha) \\ &= V_0\cos\left[k\left(\frac{\omega}{k}t - z\right) + \alpha\right] \end{aligned} \tag{20.2.1}$$

[2]The microwave community prefers β while the optics community prefers k.

We have seen before that a function $f(vt - z)$ moves with a velocity of v to the right. Thus, this sinusoidal signal moves with a velocity ω/c which is the phase velocity or

$$v_{ph} = \frac{\omega}{k} \tag{20.2.2}$$

where, for example, $k = \omega\sqrt{\mu\varepsilon}$, inside a simple coax. Hence, its phase velocity is

$$v_{ph} = 1/\sqrt{\mu\varepsilon} \tag{20.2.3}$$

But a dielectric medium can be frequency dispersive, or $\varepsilon(\omega)$ is not a constant but a function of ω as has been shown with the Drude-Lorentz-Sommerfeld model. A function of time such as $f(vt - z)$, for a fixed z, can be Fourier decomposed into a linear superposition of many time-harmonic signals. Therefore, signals with different ω's will travel with different phase velocities.

More bizarre still is when the coax is filled with a plasma medium where

$$\varepsilon = \varepsilon_0 \left(1 - \frac{{\omega_p}^2}{\omega^2}\right) \tag{20.2.4}$$

Then, $\varepsilon < \varepsilon_0$ in the above means that the phase velocity given by (20.2.3) can be larger than the velocity of light in vacuum (assuming $\mu = \mu_0$). Also, $\varepsilon = 0$ when $\omega = \omega_p$, implying that $k = 0$; then in accordance to (20.2.2), $v_{ph} = \infty$. These ludicrous observations can be justified or understood only if we can show that information can only be sent by using a wave packet.[3] The same goes for energy which can only be sent by wave packets, but not by CW signal; only in this manner can a finite amount of energy be sent. Therefore, it is prudent for us to study the velocity of a wave packet which is not a mono-chromatic signal. These wave packets can only travel at the group velocity as shall be shown, which is physical and always less than the velocity of light.

[3]In information theory, according to Shannon, the basic unit of information is a bit, which can only be sent by a digital signal, or a wave packet, or a broadband signal. A wave packet cannot be monochromatic or very narrow band.

20.2.2 Group Velocity

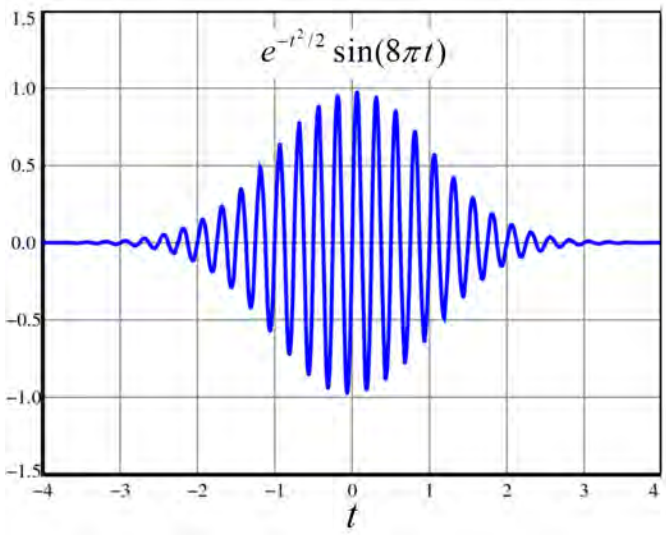


Figure 20.4: A Gaussian wave packet as a function of t for a fixed z. It can be thought of as a linear superposition of many monochromatic waves with slightly different frequencies. If one Fourier transforms the above signal, it will be a narrow-band signal centered about certain ω_0 (courtesy of Wikimedia [156]).

Now, consider a narrow-band wave packet as shown in Figure 20.4. It cannot be mono-chromatic, but can be written as a linear superposition of many frequencies. One way to express this is to write this wave packet as an integral in terms of Fourier transform, or a summation over many frequencies. Namely, at a fixed z,[4]

$$V(z,t) = \int_{-\infty}^{\infty} d\omega \underset{\sim}{V}(z,\omega) e^{j\omega t} \tag{20.2.5}$$

To make $V(z,t)$ be related to a traveling wave, we assume that $\underset{\sim}{V}(z,\omega)$ is the solution to the one-dimensional Helmholtz equation. We have encountered this equation in solving 1D Maxwell's equations in (7.2.6) or also in the transmission line theory as seen in (15.2.5) and (15.2.6). There, μ, ε, L and C can be functions of frequency and yet the equations are still valid when phasor technique is used. In general,

$$\frac{d^2}{dz^2}\underset{\sim}{V}(z,\omega) + k^2(\omega)\underset{\sim}{V}(z,\omega) = 0 \tag{20.2.6}$$

[4]The Fourier transform technique is akin to the phasor technique, but different. For simplicity, we will use $\underset{\sim}{V}(z,\omega)$ to represent the Fourier transform of $V(z,t)$.

In both Maxwell's equations and the TEM mode of a transmission line, $k^2 = \omega^2\mu\varepsilon$. The previous derivations have been for dispersionless media. But one can easily extend the derivation in Section 7.2 to a dispersive medium where $V(z,\omega) = E_x(z,\omega)$. Alternatively, one can generalize the derivation in Section 15.2 to the case of dispersive transmission lines, where, for instance, the co-axial transmission line is filled with a dispersive material. Then $k^2 = \omega^2\mu_0\varepsilon(\omega)$. Thus, upon solving the above ordinary differential equation, (20.2.6), for a fixed ω, one obtains that

$$V(z,\omega) = V_0(\omega)e^{-jkz}$$

where $k = \omega\sqrt{\mu_0\varepsilon(\omega)}$, and

$$V(z,t) = \int_{-\infty}^{\infty} d\omega V_0(\omega)e^{j(\omega t-kz)} \tag{20.2.7}$$

Here, since we have to connect a broadband time-domain signal to its Fourier transform, we will use Fourier transform technique rather than phasor technique. At this point, it is prudent to notice that if the medium in dispersionless, then $k = \omega\sqrt{\mu_0\varepsilon} = \omega/v_{ph}$ where $v_{ph} = 1/\sqrt{\mu_0\varepsilon}$ and ε is independent of frequency.[5] Then all Fourier components travel with the same phase velocity. Thus $V(z,t) = f(v_{ph}t - z)$, which is an arbitrary function traveling with velocity v_{ph} without distortion.

We shall next look at the dispersive medium case. For this medium, different Fourier components will travel with different phase velocities. If these phase velocities are all different, then the pulse travelling through it will be distorted, and become quite complicated. To simplify matters, we will assume a narrow-band pulse and find the velocity of this narrow-band pulse. As shall be shown, if the bandwidth is narrow enough, the envelope of the pulse will travel with a different velocity called the group velocity, which is different from the phase velocity. This means that the group velocity can only be defined for narrow-band pulse (also called quasi-monochromatic pulse), which is an asymptotic concept.

In the above, $V(z,t)$ is real value. As such, the negative frequency components of the above integral have to be complex conjugate of the positive frequency components. We can also rewrite the above as

$$V(z,t) = \int_{-\infty}^{0} d\omega V_0(\omega)e^{j(\omega t-kz)} + \int_{0}^{\infty} d\omega V_0(\omega)e^{j(\omega t-kz)} \tag{20.2.8}$$

Using the fact that $V_0(-\omega) = V_0^*(\omega)$ and that $k(-\omega) = k^*(\omega)$, we can write the above as sum over only the $+\omega$ part of the integral and take twice the real part of the integral.

$$V(z,t) = 2\Re e \int_{0}^{\infty} d\omega V_0(\omega)e^{j(\omega t-kz)} \tag{20.2.9}$$

In the general case, k may be a complicated function of ω as shown in Figure 20.5.

[5]There is no truly dispersionless medium except for vacuum.

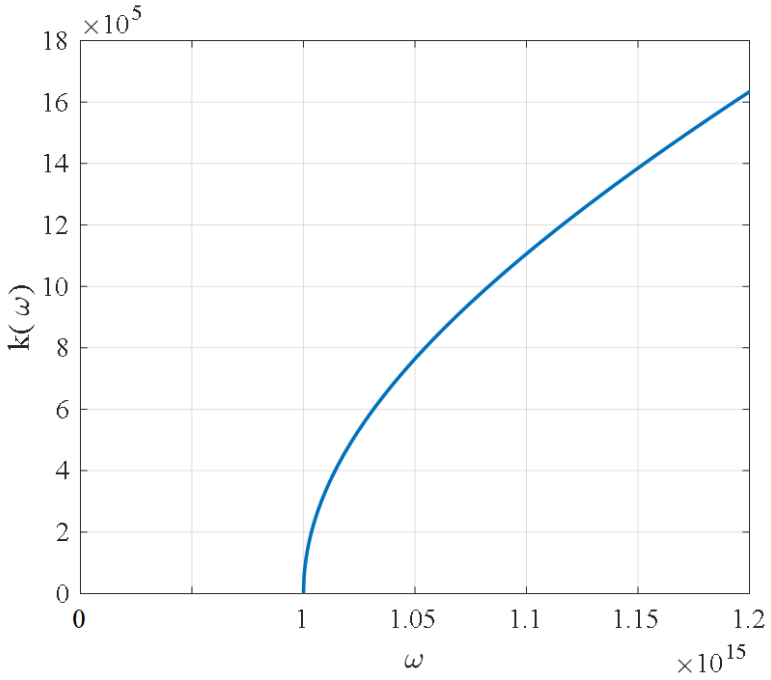


Figure 20.5: A typical frequency dependent $k(\omega)$ for a plasma medium, albeit the frequency dependence can be more complicated than shown here.

Since this is a wave packet, we assume that $V_0(\omega)$ is narrow band (quasi-monochromatic) and centered about a frequency ω_0, the carrier frequency as shown in Figure 20.6. Therefore, when the integral in (20.2.7) is performed, we need only to sum over a narrow range of frequencies in the vicinity of ω_0. Henceforth, we can approximate $k(\omega)$ in the integrand in the vicinity of $\omega = \omega_0$ by Taylor series expansion, and let

$$k(\omega) \cong k(\omega_0) + (\omega - \omega_0)\frac{dk(\omega_0)}{d\omega} + \frac{1}{2}(\omega - \omega_0)^2 \frac{d^2k(\omega_0)}{d\omega^2} + \cdots \tag{20.2.10}$$

Since we need to integrate over $\omega \approx \omega_0$, we can substitute (20.2.10) into (20.2.9) and rewrite it as

$$V(z,t) \cong 2\Re e\left[e^{j[\omega_0 t - k(\omega_0)z]} \underbrace{\int_0^\infty d\omega V_0(\omega) e^{j(\omega-\omega_0)t} e^{-j(\omega-\omega_0)\frac{dk}{d\omega}z}}_{F\left(t-\frac{dk}{d\omega}z\right)} \right] \tag{20.2.11}$$

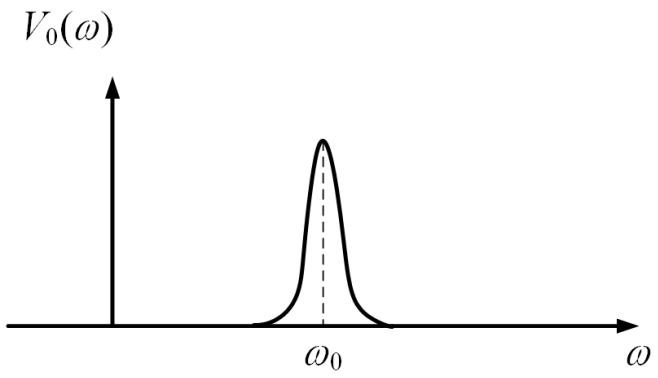


Figure 20.6: The frequency spectrum of $V_0(\omega)$ is the Fourier transform of $V(z=0,t)$ (see (20.2.7)). As can be seen, it is a narrow-band signal. Only with a narrow-band signal can we define what a group velocity is. It is the velocity of the envelope of a pulse function.

where more specifically, in the above,

$$F\left(t-\frac{dk}{d\omega}z\right)=F\left(t-z/v_g\right)=\int_0^\infty d\omega V_0(\omega)e^{j(\omega-\omega_0)t}e^{-j(\omega-\omega_0)\frac{dk}{d\omega}z} \tag{20.2.12}$$

where $v_g=\frac{d\omega}{dk}$ is the velocity of the function $F\left(t-z/v_g\right)$. Since $V_0(\omega)\neq 0$ over a small range of frequency near ω_0, (see Figure 20.6), by a change of variable by letting $\Omega=\omega-\omega_0$, it becomes

$$F\left(t-z/v_g\right)\cong\int_{-\Delta}^{+\Delta}d\Omega V_0(\Omega+\omega_0)e^{j\Omega(t-z/v_g)} \tag{20.2.13}$$

When Ω ranges from $-\Delta$ to $+\Delta$ in the above integral, the value of ω ranges from $\omega_0-\Delta$ to $\omega_0+\Delta$. It is assumed that outside this range of ω, as shown in Figure 20.6. $V_0(\omega)$ is sufficiently small so that its value can be ignored.

The above itself is a Fourier transform integral: its physical meaning is that this function $F\left(t-z/v_g\right)$ involves only the low frequencies of the Fourier spectrum where $e^{j\Omega(t-z/v_g)}$ is evaluated over small Ω values. Hence, $F\left(t-z/v_g\right)$ is a slowly varying function of its argument or $(t-z/v_g)$. Moreover, this function $F\left(t-z/v_g\right)$ moves with a velocity called the group velocity given by

$$v_g=\frac{d\omega}{dk} \tag{20.2.14}$$

Here, $F(t-\frac{z}{v_g})$ is a slowly varying envelope function or wave packet that moves with the velocity v_g as shown Figure 20.4. In (20.2.11), the envelope function $F(t-\frac{z}{v_g})$ is multiplied by the rapidly varying function

$$e^{j[\omega_0 t-k(\omega_0)z]} \tag{20.2.15}$$

before one takes the real part of the entire function. Hence, this rapidly varying part represents the rapidly varying carrier frequency shown in Figure 20.4. More importantly, this carrier, the rapidly varying part of the signal, moves with the velocity

$$v_{ph} = \frac{\omega_0}{k(\omega_0)} \tag{20.2.16}$$

which is the phase velocity. Thus we have elucidated the wave physics that a narrow-band pulse has an envelope that travels with the group velocity v_g, while the "innards" of the pulse, which include the wave inside the envelope, travels with the phase velocity v_{ph}.

20.3 Wave Guidance in a Layered Media

Now that we have understood phase and group velocity, we are at ease with studying the propagation of a guided wave in a layered medium. We have seen that in the case of a surface plasmonic resonance, the value of β_x, given by (19.1.10), and reproduced here as:

$$\beta_x = \omega\sqrt{\mu}\sqrt{\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1+\varepsilon_2}} \tag{20.3.1}$$

If β_x is linearly proportional to ω, it corresponds to the dispersionless case. Then a pulse, as previously discussed, will not be distorted as all Fourier components travel with the same velocty. However, if medium 2 is a plasma medium, ε_2 can be a strong function of frequency, and the group velocity is very different from the phase velocity.

The surface plasmon polariton wave is guided by an interface because the Fresnel reflection coefficient becomes infinite. This physically means that a reflected wave exists even if an incident wave is absent or vanishingly small. This condition can be used to find a guided mode in a layered medium, namely, to find the condition under which the composite reflection coefficient (20.1.2) becomes infinite.[6]

20.3.1 Transverse Resonance Condition

Therefore, to have a guided mode exist in a layered medium due to multiple bounces, the composite reflection coefficient becomes infinite, the denominator of (20.1.2) is zero, or that

$$1 + R_{12}\tilde{R}_{23}e^{-2j\beta_{2z}l_2} = 0 \tag{20.3.2}$$

where l_2 is the thickness of the dielectric slab. Since $R_{12} = -R_{21}$, the above can be written as

$$1 = R_{21}\tilde{R}_{23}e^{-2j\beta_{2z}l_2} \tag{20.3.3}$$

The above has the physical meaning that the wave, after going through two reflections at the two interfaces, the 21, and 23 interfaces, which are R_{21} and $\tilde{R}_{23}$, respectively, plus a phase delay given

[6] As mentioned previously in Section 19.1.1, this is equivalent to finding a solution to a problem with no driving term (forcing function), or finding the homogeneous solution to an ordinary differential equation or partial differential equation. It is also equivalent to finding the null space solution of a matrix equation.

by $e^{-2j\beta_{2z}l_2}$, becomes itself again. This is also known as the transverse resonance condition. When specialized to the case of a dielectric slab with two interfaces and three regions, the above becomes

$$1 = R_{21}R_{23}e^{-2j\beta_{2z}l_2} \tag{20.3.4}$$

The above can be generalized to finding the guided mode in a general layered medium. It can also be specialized to finding the guided mode of a dielectric slab.

Exercises for Lecture 20

Problem 20-1: This problem is about physical interpretation of the composite reflection coefficient for a layered medium.

(i) By using $T_{ij} = 1 + R_{ij}$ and that $R_{ij} = -R_{ji}$, show that the composite reflection coefficient can be written as

$$\tilde{R}_{12} = \frac{R_{12} + \tilde{R}_{23} e^{-2j\beta_{2z} l_2}}{1 + R_{12}\tilde{R}_{23} e^{-2j\beta_{2z} l_2}} = R_{12} + \frac{T_{12} T_{21} R_{23} e^{-2j\beta_{2z} l_2}}{1 + R_{12} R_{23} e^{-2j\beta_{2z} l_2}}$$

where l_2 is the thickness of region 2.

(ii) Use the geometric series expansion that $1/(1-x) = 1 + x + x^2 + x^3 + \cdots$, show that the composite reflection coefficient can be rewritten as

$$\tilde{R}_{12} = R_{12} + T_{12} R_{23} T_{21} e^{-2j\beta_{2z} l_2} + T_{12} R_{23}^2 R_{21} T_{21} e^{-4j\beta_{2z} l_2} + \cdots .$$

(iii) Give the physical meanings of each of the terms above, and the meaning of the phase delay term.

(iv) Explain what phase and group velocities are.

(v) Derive equation (20.2.11) of the lecture notes and explain why $F\left(t - \frac{dk}{d\omega} z\right)$ is a slowly varying function, and explain what a group velocity is.

Chapter 21

Dielectric Slab Waveguides

As mentioned before, the dielectric slab waveguide shares many salient features with the optical fiber waveguide, one of the most important waveguides of this century.[1] The analysis of the optical fiber requires the Maxwellian solution in cylindrical coordinates which is beyond our scope, but it can be found in [1, 90, 157]. Before we embark on studying dielectric slab waveguides, we will revisit the transverse resonance again. The transverse resonance condition allows one to derive the guidance conditions for a dielectric slab waveguide easily without having to match the boundary conditions at the interfaces again: The boundary conditions are already embedded in the derivation of the Fresnel reflection coefficients. Much of the materials in this lecture can also be found in [90, 47, 34].

21.1 Generalized Transverse Resonance Condition

The generalized transverse resonance condition is a powerful condition that can be used to derive the guidance conditions of modes in a layered medium. To derive this condition, we first have to realize that a guided mode in a waveguide is due to the coherent or constructive interference of the waves. This implies that if a plane wave starts at position 1 (see Figure 21.1)[2] and is multiply reflected as shown, it will regain its original phase in the x direction at position 5. Since this mode progresses in the z direction, all these waves (also known as partial waves) are in phase in the z direction by the phase matching condition. Otherwise, the boundary conditions cannot be satisfied. (That is, waves at 1 and 5 will gain the same phase in the z direction.) But, for them to add coherently or interfere coherently in the x direction, the transverse phase in the x direction at 5 must be the same as 1.

Assuming that the wave starts with amplitude 1 at position 1, that the distance between 1 and 2 is t, then it will gain a transverse phase of $e^{-j\beta_{0x}t}$ when it reaches position 2. Upon reflection at $x = x_2$, at position 3, the wave becomes $\tilde{R}_+ e^{-j\beta_{0x}t}$ where $\tilde{R}_+$ is the generalized reflection coefficient

[1] The optical fiber has also very low loss of the order of 0.1 dB/km.

[2] The waveguide convention is to assume the direction of propagation of a mode to be in the z direction. Since we are analyzing a guided mode in a layered medium, z axis is as shown in this figure, which is parallel to the interfaces. This is different from before.

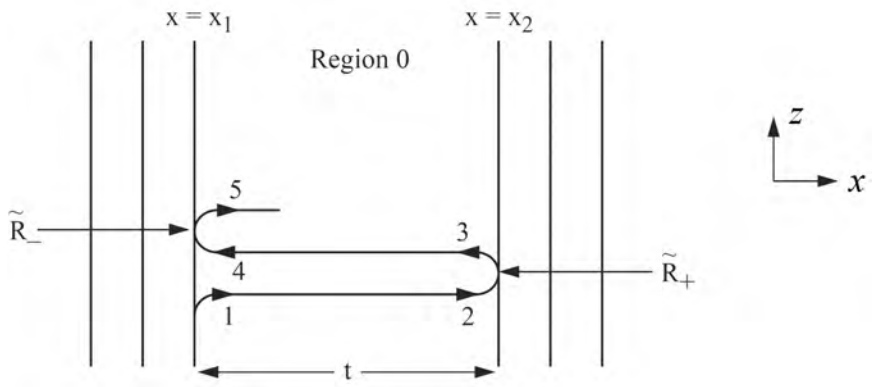


Figure 21.1: The transverse resonance condition for a layered medium. For constructive interference to occur, the phase of the wave at position 5 should be equal to the transverse phase at position 1. The above is a cartoon picture as the ray does not bend as shown.

at the right interface of Region 0. Finally, at position 5, it becomes $\tilde{R}_-\tilde{R}_+e^{-2j\beta_{0x}t}$ where $\tilde{R}_-$ is the generalized reflection coefficient at the left interface of Region 0. For constructive interference to occur or for the mode to exist, we require that transverse phase of the wave at position 5 is the same as that at position 1, or

$$\tilde{R}_-\tilde{R}_+e^{-2j\beta_{0x}t} = 1 \tag{21.1.1}$$

The above is the generalized transverse resonance condition for the guidance condition for a plane wave mode traveling in a layered medium.

Alternatively, we can look at the generalized reflection coefficient of the previous lecture. Internal modes exist in the layered medium when the generalized reflection coefficient goes to infinity. For a three-layered medium, looking at (20.3.3), the denominator of the equation will be zero when

$$R_{21}R_{23}e^{-2j\beta_{2x}t} = 1 \tag{21.1.2}$$

Here we assume the wave number in the middle region is β_{2x} for a wave propagating in the x direction.

21.1.1 Parallel Plate Waveguide

Note that for equation (21.1.1), when we have two parallel metallic plates, where the metallic plates are assumed to be PEC, then $R_\pm^{TM} = 1$, and $R_\pm^{TE} = -1$,[3] and the guidance condition becomes

$$1 = e^{-2j\beta_{0x}t} \quad \Rightarrow \beta_{0x} = \frac{m\pi}{t}, \qquad m = 0, 1, 2, \ldots, \tag{21.1.3}$$

These are just the guidance conditions for parallel plate waveguides. In this waveguide, the modes are guided by total reflections at the air-metallic interface due to the impenetrability of a PEC

[3]This can be seen from the Fresnel reflection coefficient by setting one of the media to be PEC.

surface. Details of this waveguide are given in the ECE 350X lecture notes as well as in Kong [47, 34].

21.2 Dielectric Slab Waveguide

The most important dielectric waveguide of the modern world is the optical fiber, whose invention was credited to Charles Kao [142]. He was awarded the Nobel prize in 2009 [158]. However, the closed-form analysis of the optical fiber requires the use of cylindrical coordinates and special functions such as Bessel functions. In order to capture the essence of the optical fiber or different dielectric waveguides without the use of special functions such as Bessel functions, one can study the slab dielectric waveguide, which shares many salient wave-physics features. We start with analyzing the TE modes in this waveguide. (This waveguide is also used as thin-film optical waveguides (see Figure 21.2).)

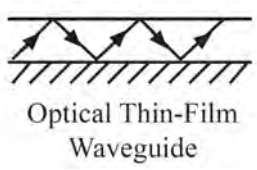


Figure 21.2: An optical thin-film waveguide is made by coating a thin dielectric film or sheet on a metallic surface. The wave is guided by total internal reflection at the top interface, and by metallic reflection at the bottom interface, where the wave is totally reflected as well.

21.2.1 TE Case

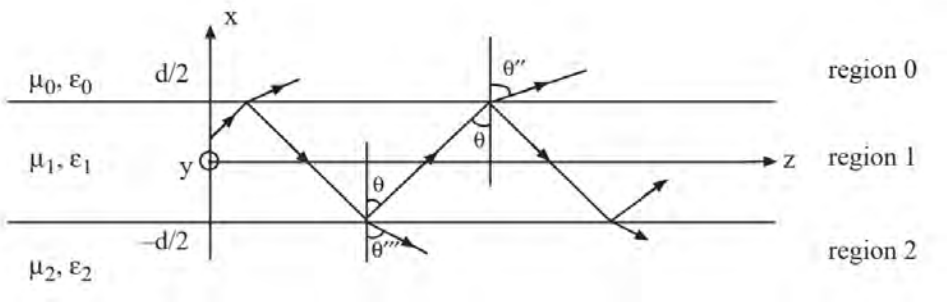


Figure 21.3: Schematic for the analysis of a guided mode in the dielectric waveguide. Total internal reflections occur at the top and bottom interfaces. If the waves add coherently, the wave is guided along the dielectric slab.

We shall look at the application of the transverse resonance condition to a TE wave guided in a dielectric slab waveguide. Again, in accordance with convention, we assume the direction of propagation of the guided mode to be in the z direction. Specializing the above equation to the dielectric slab waveguide shown in Figure 21.3, we have the guidance condition as

$$1 = R_{10}R_{12}e^{-2j\beta_{1x}d} \tag{21.2.1}$$

where d is the thickness of the dielectric slab. Guidance of a mode is due to total internal reflection, and hence, we expect Region 1 to be optically more dense (in terms of optical refractive indices) than Regions 0 and 2.[4]

To simplify the analysis further, we assume Region 2 to be the same as Region 0 so that $R_{12} = R_{10}$. The new guidance condition is then

$$1 = R_{10}^2e^{-2j\beta_{1x}d} \tag{21.2.2}$$

In the above,

$$R_{10} = \frac{\mu_0\beta_{1x} - \mu_1\beta_{0x}}{\mu_0\beta_{1x} + \mu_1\beta_{0x}} \tag{21.2.3}$$

By phase-matching, β_z is the same in all the three regions of Figure 21.3. In the above, $\beta_{ix} = \sqrt{\beta_i^2 - \beta_z^2}$, by expressing all the β_{ix} in terms of the variable β_z, the above (21.2.2) becomes an implicit equation to be solved for β_z. Also, we assume that $\varepsilon_1 > \varepsilon_0$ so that total internal reflection occurs at both interfaces as the wave bounces around so that $\beta_{0x} = \beta_{2x} = -j\alpha_{0x}$. In other words, for TE polarization, the local, single-interface, or Fresnel reflection coefficient is

$$R_{10} = \frac{\mu_0\beta_{1x} - \mu_1\beta_{0x}}{\mu_0\beta_{1x} + \mu_1\beta_{0x}} = \frac{\mu_0\beta_{1x} + j\mu_1\alpha_{0x}}{\mu_0\beta_{1x} - j\mu_1\alpha_{0x}} = e^{j\theta_{TE}} \tag{21.2.4}$$

where θ_{TE} is the Goos-Hanschen shift for total internal reflection. It is given by

$$\theta_{TE} = 2\tan^{-1}\left(\frac{\mu_1\alpha_{0x}}{\mu_0\beta_{1x}}\right) \tag{21.2.5}$$

The guidance condition for constructive interference according to (21.2.1) and (21.2.2), after using (21.2.4) and (21.2.5), is such that

$$2\theta_{TE} = 2\beta_{1x}d + 2l\pi \tag{21.2.6}$$

From the above, in view of (21.2.5), we divide the above by four, and taking its tangent, we get

$$\tan\left(\frac{\theta_{TE}}{2}\right) = \tan\left(\frac{l\pi}{2} + \frac{\beta_{1x}d}{2}\right) \tag{21.2.7}$$

or using (21.2.5) for the left-hand side,

$$\frac{\mu_1\alpha_{0x}}{\mu_0\beta_{1x}} = \tan\left(\frac{l\pi}{2} + \frac{\beta_{1x}d}{2}\right) \tag{21.2.8}$$

[4]Optically more dense means higher optical refractive index, or higher dielectric constant. As a reminder, the refractive index of a medium with permittivity ε and permeability μ is $n = \sqrt{\mu\varepsilon/(\mu_0\varepsilon_0)}$ where ε_0 and μ_0 are the permittivity and permeability, respectively, of vacuum.

The above gives rise to

$$\mu_1 \alpha_{0x} = \mu_0 \beta_{1x} \tan\left(\frac{\beta_{1x} d}{2}\right), \qquad l \text{ even} \tag{21.2.9}$$

$$-\mu_1 \alpha_{0x} = \mu_0 \beta_{1x} \cot\left(\frac{\beta_{1x} d}{2}\right), \qquad l \text{ odd} \tag{21.2.10}$$

It can be shown that when l is even, the mode profile is even, whereas when l is odd, the mode profile is odd. The above can also be rewritten as

$$\frac{\mu_0}{\mu_1} \frac{\beta_{1x} d}{2} \tan\left(\frac{\beta_{1x} d}{2}\right) = \frac{\alpha_{0x} d}{2}, \qquad \text{even modes} \tag{21.2.11}$$

$$-\frac{\mu_0}{\mu_1} \frac{\beta_{1x} d}{2} \cot\left(\frac{\beta_{1x} d}{2}\right) = \frac{\alpha_{0x} d}{2}, \qquad \text{odd modes} \tag{21.2.12}$$

Again, the above equations can be expressed in the β_z variable, but they do not have closed form solutions, except for graphical solutions (or numerical solutions). We shall discuss their graphical solutions next.[5]

To solve the above graphically, it is best to plot them in terms of one common variable. It turns out the β_{1x} is the simplest common variable to use for graphical solutions since the left-hand side of (21.2.11) and (21.2.12) are simple functions of β_{1x}. To this end, using the fact that

$$-\alpha_{0x}^2 = \beta_0^2 - \beta_z^2$$

, and that

$$\beta_{1x}^2 = \beta_1^2 - \beta_z^2$$

, eliminating β_z from these two equations by subtraction, one can show that α_{0x} on the right-hand side becomes

$$\alpha_{0x} = [\omega^2(\mu_1 \epsilon_1 - \mu_0 \epsilon_0) - \beta_{1x}^2]^{\frac{1}{2}} \tag{21.2.13}$$

By regarding β_{1x} as the abscissa, and α_{0x} as the ordinate, we notice that the above is the equation of a circle. Thus the right-hand side of (21.2.11) and (21.2.12) can be simplified and become

$$\begin{aligned} \frac{\mu_0}{\mu_1} \frac{\beta_{1x} d}{2} \tan\left(\frac{\beta_{1x} d}{2}\right) &= \frac{\alpha_{0x} d}{2} \\ &= \sqrt{\omega^2(\mu_1 \epsilon_1 - \mu_0 \epsilon_0)\frac{d^2}{4} - \left(\frac{\beta_{1x} d}{2}\right)^2}, \qquad \text{even modes} \end{aligned} \tag{21.2.14}$$

$$\begin{aligned} -\frac{\mu_0}{\mu_1} \frac{\beta_{1x} d}{2} \cot\left(\frac{\beta_{1x} d}{2}\right) &= \frac{\alpha_{0x} d}{2} \\ &= \sqrt{\omega^2(\mu_1 \epsilon_1 - \mu_0 \epsilon_0)\frac{d^2}{4} - \left(\frac{\beta_{1x} d}{2}\right)^2}, \qquad \text{odd modes} \end{aligned} \tag{21.2.15}$$

[5]This technique has been put together by the community of scholars in the optical waveguide area.

We can solve the above graphically by plotting

$$y_1 = \frac{\mu_0}{\mu_1}\frac{\beta_{1x}d}{2}\tan\left(\frac{\beta_{1x}d}{2}\right), \quad \text{even modes} \tag{21.2.16}$$

$$y_2 = -\frac{\mu_0}{\mu_1}\frac{\beta_{1x}d}{2}\cot\left(\beta_{1x}\frac{d}{2}\right), \quad \text{odd modes} \tag{21.2.17}$$

$$y_3 = \left[\omega^2(\mu_1\epsilon_1 - \mu_0\epsilon_0)\frac{d^2}{4} - \left(\frac{\beta_{1x}d}{2}\right)^2\right]^{\frac{1}{2}} = \frac{\alpha_{0x}d}{2} \tag{21.2.18}$$

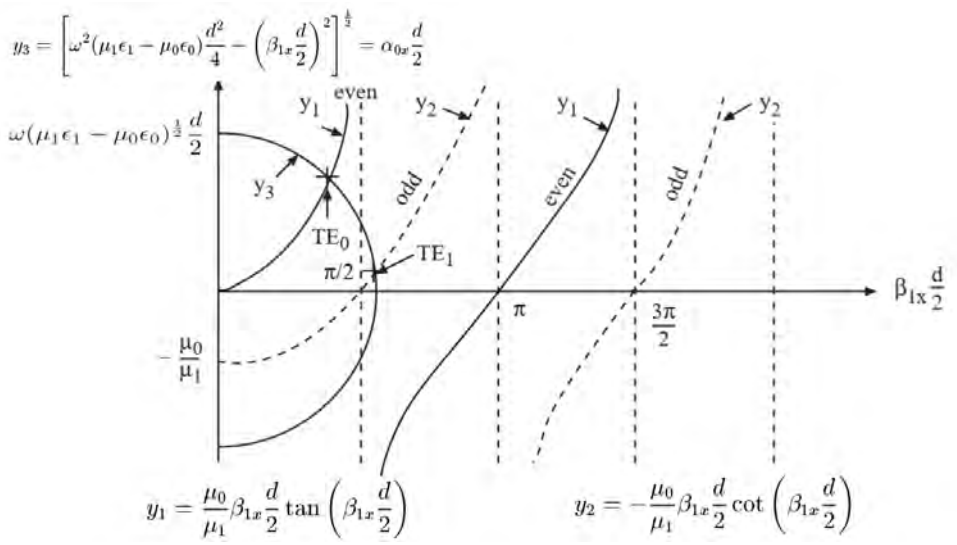


Figure 21.4: A way to solve (21.2.14) and (21.2.15) is via a graphical method. In this method, both the right-hand side and the left-hand side of the equations are plotted on the same plot. The solutions are at the intersection points of these plots.

In the above, y_3 is the equation of a circle in terms of $\beta_{1x}\frac{d}{2}$. The radius of the circle is given by

$$\omega(\mu_1\epsilon_1 - \mu_0\epsilon_0)^{\frac{1}{2}}\frac{d}{2}. \tag{21.2.19}$$

The solutions to (21.2.14) and (21.2.15) are given by the intersections of y_3 with y_1 and y_2. To increase the number of solutions, we can increase the radius of the circle defined by y_3. We note from (21.2.1) that the radius of the circle can be increased in three ways: (i) by increasing the frequency ω, (ii) by increasing the contrast $\frac{\mu_1\epsilon_1}{\mu_0\epsilon_0}$, and (iii) by increasing the thickness d of the slab.[6]

[6]These are important salient features of a dielectric waveguide. These features are also shared by the optical fiber.

By increasing these three parameters, then the number of trapped modes (or guided modes) inside the slab waveguide increases. The mode profiles of the first two modes are shown in Figure 21.5.

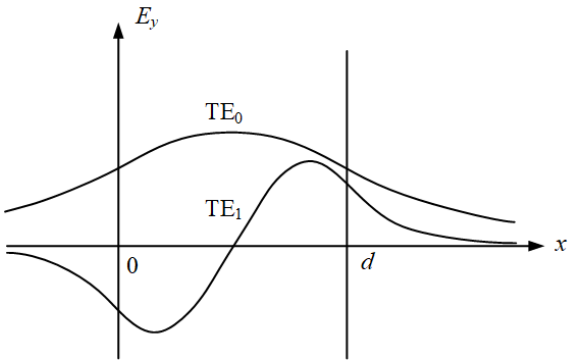


Figure 21.5: Mode profiles of the TE$_0$ and TE$_1$ modes of a dielectric slab waveguide (courtesy of J.A. Kong [34]).

When $\beta_{0x} = -j\alpha_{0x}$, the reflection coefficient for total internal reflection is

$$R_{10}^{TE} = \frac{\mu_0\beta_{1x} + j\mu_1\alpha_{0x}}{\mu_0\beta_{1x} - j\mu_1\alpha_{0x}} = \exp\left[+2j\tan^{-1}\left(\frac{\mu_1\alpha_{0x}}{\mu_0\beta_{1x}}\right)\right] \tag{21.2.20}$$

and $\left|R_{10}^{TE}\right| = 1$. Hence, the wave is guided by total internal reflections at the two interfaces.

Cut-off happens when the total internal reflection ceases to occur, i.e. when the frequency decreases such that $\alpha_{0x} = 0$. When this happens, the mode is not trapped or confined anymore or the wave is not evanescent outside the slab. From Figure 21.4, we see that $\alpha_{0x} = 0$ when

$$\omega(\mu_1\epsilon_1 - \mu_0\epsilon_0)^{\frac{1}{2}}\frac{d}{2} = \frac{m\pi}{2}, \qquad m = 0, 1, 2, 3, \ldots \tag{21.2.21}$$

or

$$\omega_{mc} = \frac{m\pi}{d(\mu_1\epsilon_1 - \mu_0\epsilon_0)^{\frac{1}{2}}}, \qquad m = 0, 1, 2, 3, \ldots \tag{21.2.22}$$

where ω_{mc} is the cutoff frequency of the m-th mode. The mode that corresponds to the m-th cut-off frequency above is labeled the TE$_m$ mode. Thus TE$_0$ mode is the mode that has no cut-off or propagates at all frequencies. This is shown in Figure 21.6 where the TE mode profiles are similar since they are dual to each other. The boundary conditions at the dielectric interface is that the tangential E and H fields have to be continuous. The TE$_0$ or TM$_0$ mode can satisfy this boundary condition at all frequencies. At the cut-off frequency, the field outside the slab is not evanescent, but has to become flat implying the $\alpha_{0x} = 0$. Therefore, the mode is not confined to the vicinity of the waveguide, and there is no guidance.

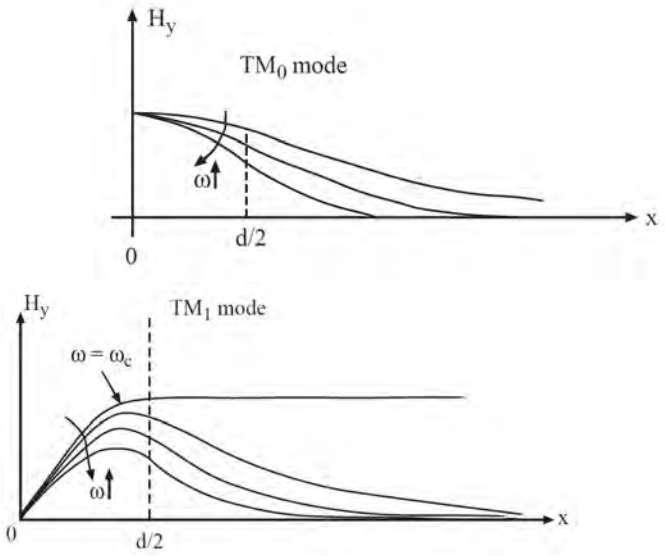


Figure 21.6: Mode profiles of the TM modes of a dielectric slab. The TE modes are dual to the TM modes and have similar mode profiles.

Next, we will elucidate more physics of the dielectric slab guided mode. At cut-off, $\alpha_{0x} = 0$, and from the dispersion relation that $\alpha_{0x}^2 = \beta_z^2 - \beta_0^2$, we conclude that

$$\beta_z = \beta_0 = \omega\sqrt{\mu_0\epsilon_0},$$

for all the modes. Hence, the phase velocity, ω/β_z, and the group velocity, $d\omega/d\beta_z$ are those of the outer region. This is because when $\alpha_{0x} = 0$, the wave is not evanescent outside, and the energy of the mode is predominantly carried by the exterior field.

When ω becomes increasingly larger, the radius of the circle in the plot of y_3 also becomes increasingly larger. As seen from Figure 21.4, the solution for $\beta_{1x} \to \frac{l\pi}{d}$ for all the modes. In other words, the semi-circle in Figure 21.4 will intersect all the vertical dotted lines as ω increases.

From the dispersion relation for Region 1, since $\omega^2\mu_1\epsilon_1 \gg \beta_{1x}^2 \approx (l\pi/d)^2$, and we have

$$\beta_z = \sqrt{\omega^2\mu_1\epsilon_1 - \beta_{1x}^2} \approx \sqrt{\omega^2\mu_1\epsilon_1 - (l\pi/d)^2} \approx \omega\sqrt{\mu_1\epsilon_1}, \qquad \omega \to \infty \tag{21.2.23}$$

Hence the group and phase velocities approach that of the dielectric slab. This is because when $\omega \to \infty$, $\alpha_{0x} \to \infty$, implying the rapid exponential decay of the fields outside the waveguide. Therefore, the fields are trapped or confined in the slab and propagating within it. Because of this, the dispersion diagram of the different modes appear as shown in Figure 21.7.

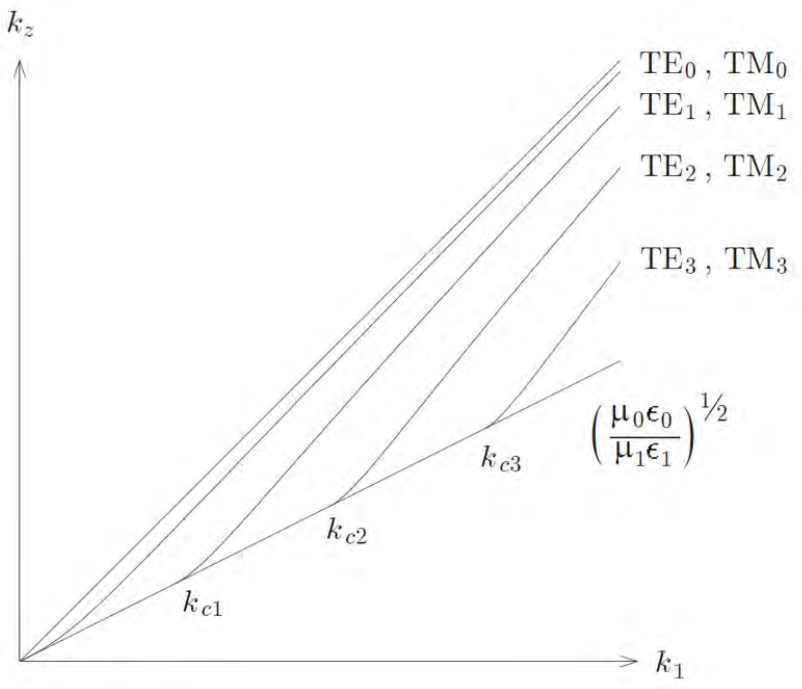


Figure 21.7: Here, we have k_z versus k_1 plots for dielectric slab waveguide. Near its cut-off, the energy of the mode is in the outer region, and hence, its group velocity given by $\frac{d\omega}{dk_z}$ is close to that of the outer region. At high frequencies, the mode is tightly bound to the slab, and its group velocity approaches that of the dielectric slab (courtesy of J.A. Kong [34]).

In Figure 21.7,[7] k_{c1}, k_{c2}, and k_{c3} are the cut-off wave number or frequency of the first three modes. In the same figure, if the wave is in free space, then $k_z = \omega\sqrt{\mu_0\varepsilon_0}$. This is also called the light line. In the $k_z - k_1$ plot, we see that

$$k_z = \left(\frac{\mu_0\varepsilon_0}{\mu_1\varepsilon_1}\right)^{1/2} k_1 \tag{21.2.24}$$

Therefore, the slope of the line is $\left(\frac{\mu_0\varepsilon_0}{\mu_1\varepsilon_1}\right)^{1/2}$. When the mode is at cut-off, the energy of the mode is in the outer region and the dispersion curve hugs the light line (so that they have the same slope), so that both the guided mode and the free space have the same group velocity. As the frequency increases above the cut-off, the mode is better trapped or confined to the waveguide region. In this case, the group velocity of the mode approaches that of the waveguide medium.

[7]Please note again that in this course, we will use β and k interchangeably for wavenumbers.

21.2.2 TM Case

For the TM case, a similar guidance condition analogous to (21.2.1) can be derived using duality principle, and with the understanding that the reflection coefficients in (21.2.1) are now TM reflection coefficients. Similar derivations show that the above guidance conditions, for $\epsilon_2 = \epsilon_0$, $\mu_2 = \mu_0$, reduce to

$$\frac{\epsilon_0}{\epsilon_1}\beta_{1x}\frac{d}{2}\tan\beta_{1x}\frac{d}{2} = \sqrt{\omega^2(\mu_1\epsilon_1 - \mu_0\epsilon_0)\frac{d^2}{4} - \left(\beta_{1x}\frac{d}{2}\right)^2}, \qquad \text{even modes} \tag{21.2.25}$$

$$-\frac{\epsilon_0}{\epsilon_1}\beta_{1x}\frac{d}{2}\cot\beta_{1x}\frac{d}{2} = \sqrt{\omega^2(\mu_1\epsilon_1 - \mu_0\epsilon_0)\frac{d^2}{4} - \left(\beta_{1x}\frac{d}{2}\right)^2}, \qquad \text{odd modes} \tag{21.2.26}$$

21.2.3 A Note on Cut-Off of Dielectric Waveguides

The concept of cut-off in dielectric waveguides is quite different from that of hollow waveguides that we shall learn next. A mode is guided in a dielectric waveguide if the wave is trapped (or confined) inside the dielectric slab. The trapping is due to the total internal reflections at the top and the bottom interfaces of the waveguide. When total internal reflection ceases to occur at any of the two interfaces, the wave is not guided or trapped inside the dielectric slab anymore. This happens when $\alpha_{ix} = 0$ where i can indicate the top-most or the bottom-most region. In other words, the wave ceases to be evanescent in either the top-most or the bottom-most region.

21.2.4 Alternative Derivation of the Guidance Condition

The previous derivation does not allow us to determine the mode profiles as in Figure 21.5. It would not let us know that if l is even, the mode profile is even and so on. To get this information, we need an alternative derivation as is shown here. We focus on the TE case, because the TM case is easily arrived by using duality principle. For the TE case, E_y is a solution to the wave equation in each region. In Region 0, we assume a solution of the form

$$E_{0y} = E_0 e^{-j\beta_{0x}x - j\beta_z z} \tag{21.2.27}$$

that corresponds to an upgoing travelling wave from the interface, and

$$\beta_{0x}^2 + \beta_z^2 = \omega^2\mu_0\epsilon_0 = \beta_0^2 \tag{21.2.28}$$

In Region 1, we assume a two-way travelling wave and solution is of the form

$$E_{1y} = [A_1 e^{-j\beta_{1x}x} + B_1 e^{j\beta_{1x}x}]e^{-j\beta_z z} \tag{21.2.29}$$

where the first term is an upgoing wave while the second term is a downgoing wave, and

$$\beta_{1x}^2 + \beta_z^2 = \omega^2\mu_1\epsilon_1 = \beta_1^2 \tag{21.2.30}$$

In Region 2, the solution is a downgoing travelling wave of the form

$$E_{2y} = E_2 e^{j\beta_{2x}x - j\beta_z z} \tag{21.2.31}$$

with

$$\beta_{2x}^2 + \beta_z^2 = \omega^2 \mu_2 \epsilon_2 = \beta_2^2 \tag{21.2.32}$$

We assume that all the solutions in the three regions to have the same z-variation of $e^{-j\beta_z z}$ or the same β_z by the **phase matching** condition.

In Region 1, we have an upgoing wave as well as a downgoing wave. The two waves have to be related by the reflection coefficient R^{TE} for the electric field at the boundaries. Therefore at $x = \frac{d}{2}$, an downgoing wave is a consequence of the reflection of the upgoing wave. Therefore, we have

$$B_1 e^{j\beta_{1x}\frac{d}{2}} = R_{10}^{TE} A_1 e^{-j\beta_{1x}\frac{d}{2}} \tag{21.2.33}$$

where R_{10}^{TE} is the TE reflection coefficient at the Regions 1 and 0 interface.

At $x = -\frac{d}{2}$, an upgoing wave is a consequence of the reflection of the downgoing wave. Or we have

$$A_1 e^{j\beta_{1x}\frac{d}{2}} = R_{12}^{TE} B_1 e^{-j\beta_{1x}\frac{d}{2}} \tag{21.2.34}$$

where R_{12}^{TE} is the reflection coefficient at the Regions 1 and 2 interface. Multiplying equations (21.2.33) and (21.2.33) together, we have,

$$A_1 B_1 e^{j\beta_{1x}d} = R_{12}^{TE} R_{10}^{T} E A_1 B_1 e^{-j\beta_{1x}d} \tag{21.2.35}$$

A_1 and B_1 are non-zero only if

$$1 = R_{12}^{TE} R_{10}^{TE} e^{-2j\beta_{1x}d} \tag{21.2.36}$$

The above is exactly the **guidance condition** of a dielectric slab waveguide previously derived from the transverse resonance condition. If medium 3 is equal to medium 1, then $R_{12}^{TE} = R_{10}^{TE}$, and the guidance condition becomes

$$1 = (R_{10}^{TE})^2 e^{-2j\beta_{1x}d} \tag{21.2.37}$$

Taking the square root of (21.2.37), we have

$$R_{10}^{TE} e^{-j\beta_{1x}d} = \pm 1 \tag{21.2.38}$$

When we choose the plus sign, $B_1 = A_1$ from (21.2.33), and from (21.2.29)

$$E_{1y} = 2A_1 \cos(\beta_{1x} x) e^{-j\beta_z z} \quad \Rightarrow \text{even in } x \tag{21.2.39}$$

When we choose the minus sign in (11) we have $B_1 = -A_1$, and

$$E_{1y} = -2jA_1 \sin(\beta_{1x} x) e^{-j\beta_z z} \quad \Rightarrow \text{odd in } x \tag{21.2.40}$$

It can be shown that the two solutions given by (21.2.38) map to the even and odd solutions in (21.2.11) and (21.2.12), and hence, their names.

Exercises for Lecture 21

Problem 21-1:

(i) Explain what a general transverse resonance condition is.

(ii) By going through the lecture notes on dielectric waveguides in the alternate derivation part, explain why when n in Lecture 21 (of lecture book), eqns. (21.2.9) and (21.2.10) is even or odd, they correspond to mode profiles that are even or odd.

(iii) Explain why the TE_0 mode of a dielectric slab waveguide has no cut-off frequency.

(iv) Show that the two solutions given by (21.2.38) map to the even and odd solutions in (21.2.11) and (21.2.12).

Chapter 22

Hollow Waveguides

Hollow waveguides are useful for high-power microwaves. Air has a higher breakdown voltage compared to most materials, and hence, it could be a good medium for propagating high electromagnetic energy and it has low loss. Also, hollow metallic waveguides are sufficiently shielded from the rest of the world so that interference from other sources is minimized. Furthermore, for radio astronomy, they can provide a low-noise system immune to interference. Air generally has less loss than materials, and loss is often the source of thermal noise. Therefore, a low loss waveguide is also a low noise waveguide.[1]

22.1 General Information on Hollow Waveguides

Many waveguide problems can be solved in closed form. An example is the coaxial waveguide previously discussed. In addition, there are many other waveguide problems that have closed form solutions. Closed form solutions to Laplace and Helmholtz equations are obtained by the separation-of-variables method. The separation-of-variables method works only for separable coordinate systems. (There are 11 separable coordinates for Helmholtz equation, but 13 for Laplace equation.) Some examples of separable coordinate systems are cartesian, cylindrical, and spherical coordinates. These three coordinates are about all we need to know for solving many engineering problems. For other than these three coordinates, complex special functions need to be defined for their solutions, which are hard to compute. Therefore, more complicated cases are now handled with numerical methods using computers.

When a waveguide has a center conductor or two conductors like a coaxial cable, it can support a TEM wave (or TEM mode) where both the electric field and the magnetic field are orthogonal to the direction of propagation. The uniform plane wave is an example of a TEM wave (see Section 7.3, (7.3.11)). However, when the waveguide is hollow or is filled completely with a homogeneous medium, and without a center conductor, it cannot support a TEM mode as we shall prove next.

[1] The fluctuation dissipation theorem [159, 160] says that when a system loses energy to the environment, it also receives the same amount of energy from the environment for energy conservation. In a word, a lossy system loses energy to its environment, but it also receives energy back from the environment in terms of thermal noise. Thus, the lossier a system is, the more thermal noise is needed for energy balance.

Much of the materials of this lecture can be found in [131, 34, 90].

22.1.1 Absence of TEM Mode in a Hollow Waveguide

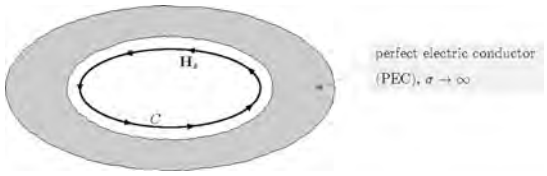


Figure 22.1: Absence of TEM mode in a hollow waveguide enclosed by a PEC wall. The magnetic field lines form a closed loop inside the waveguide due to the absence of magnetic charges.

We would like to prove by contradiction (*reductio ad absurdum*) that a hollow waveguide as shown in Figure 22.1 (i.e. without a center conductor) cannot support a TEM mode. The proof is as follows: If we assume that TEM mode does exist, then the magnetic field has to end on itself due to the absence of magnetic charges on the waveguide wall. In this case, it is clear that $\oint_C \mathbf{H}_s \cdot d\mathbf{l} \neq 0$ about any closed contour following the magnetic field lines. But Ampere's law states that the above is equal to

$$\oint_C \mathbf{H}_s \cdot d\mathbf{l} = j\omega\varepsilon \int_S \mathbf{E} \cdot d\mathbf{S} + \int_S \mathbf{J} \cdot d\mathbf{S} \tag{22.1.1}$$

The left-hand side of the above equation is clearly nonzero by the above argument. But for a hollow waveguide, $\mathbf{J} = 0$ and the above becomes

$$\oint_C \mathbf{H}_s \cdot d\mathbf{l} = j\omega\varepsilon \int_S \mathbf{E} \cdot d\mathbf{S} = j\omega\varepsilon \int_S \mathbf{E} \cdot \hat{n} dS \tag{22.1.2}$$

where $\hat{n} = \hat{z}$. Hence, this equation cannot be satisfied unless on the right-hand side, there are $E_z \neq 0$ component. This implies that a TEM mode where both E_z and H_z are zero in a hollow waveguide is impossible without a center conductor.

By the above argument, in a hollow waveguide filled with homogeneous medium, but only TE_z (TE to z) or TM_z (TM to z) modes can exist. For a TE_z wave (or TE wave), $E_z = 0$, $H_z \neq 0$ while for a TM_z wave (or TM wave), $H_z = 0$, $E_z \neq 0$. These classes of problems can be decomposed into two scalar problems like the layered medium case,[2] by using the pilot potential method. However, when the hollow waveguide is filled with a center conductor, the TEM mode can exist in addition to TE and TM modes. (It is to be noted that in such a hollow waveguide, the TEM mode is the degenerate case of either the TE or the TM modes in the waveguide. Hence, the subsequent analyses for TE and TM modes are also valid for the TEM mode if it does exist.)

We begin by studying some simple waveguide where closed form solutions exist to hollow waveguides, such as the rectangular waveguides. These closed form solutions offer physical insight into

[2]Looking at the TE and TM waves in the half-space reflection problem (see Figures 18.1 and 18.3), we see that for TE waves, $H_z \neq 0$, and for TM waves, $E_z \neq 0$.

the propagation of waves in such a waveguide. Another waveguide with slightly more complicated closed form solutions is the circular hollow waveguide. The solutions need to be sought in terms of Bessel functions. Another waveguide with very complicated closed form solutions is the elliptical waveguide. However, the solutions are too complicated to be considered; these days, the preferred method of solving these complicated problems is via numerical methods.

22.1.2 TE Case ($E_z = 0$, $H_z \neq 0$, $\mathbf{TE}_z$ case)

In this case, the field inside the waveguide is TE to z (or TE_z). To ensure such a TE field, we can write the $\mathbf{E}$ field as

$$\mathbf{E}(\mathbf{r}) = \nabla \times \hat{z}\Psi_h(\mathbf{r}) \tag{22.1.3}$$

By construction, equation (22.1.3) will guarantee that $E_z = 0$. Here, $\Psi_h(\mathbf{r})$ is a scalar potential and $\hat{z}$ is called the pilot vector.[3] (The subscript h is used because, as shall be shown, this scalar potential can be related to the z component of the $\mathbf{H}$ field.)

The waveguide is assumed source-free and filled with a lossless, homogeneous material. Eq. (22.1.3) also satisfies the source-free condition since, clearly, $\nabla \cdot \mathbf{E} = 0$. And hence, from Maxwell's equations that

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{22.1.4}$$

$$\nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E} \tag{22.1.5}$$

it can be shown that

$$\nabla \times \nabla \times \mathbf{E} - \omega^2\mu\varepsilon\mathbf{E} = 0 \tag{22.1.6}$$

Furthermore, using the appropriate vector identiy, such as the BAC-CAB formula, and using $\nabla \cdot \mathbf{E} = 0$ condition, it can be shown that the electric field $\mathbf{E}(\mathbf{r})$ satisfies the following Helmholtz wave equation (or partial differential equation) that

$$(\nabla^2 + \beta^2)\mathbf{E}(\mathbf{r}) = 0 \tag{22.1.7}$$

where $\beta^2 = \omega^2\mu\varepsilon$. Substituting (22.1.3) into (22.1.7), we get

$$(\nabla^2 + \beta^2)\nabla \times \hat{z}\Psi_h(\mathbf{r}) = 0 \tag{22.1.8}$$

In the above, we can show that $\nabla^2\nabla \times \hat{z}\Psi = \nabla \times \hat{z}(\nabla^2\Psi)$, or that these operators commute.[4] Then it follows that

$$\nabla \times \hat{z}(\nabla^2 + \beta^2)\Psi_h(\mathbf{r}) = 0 \tag{22.1.9}$$

Thus, if $\Psi_h(\mathbf{r})$ satisfies the following Helmholtz wave equation or partial differential equation

$$(\nabla^2 + \beta^2)\Psi_h(\mathbf{r}) = 0 \tag{22.1.10}$$

[3] It "pilots" the field so that it is transverse to z.

[4] This is a mathematical parlance, and a commutator is defined to be $[A, B] = AB - BA$ for two operators A and B. If these two operators commute, then $[A, B] = 0$.

then (22.1.9) is satisfied, and so is (22.1.7).[5] Hence, the $\mathbf{E}$ field constructed with (22.1.3) satisfies Maxwell's equations, if $\Psi_h(\mathbf{r})$ satisfies (22.1.10).

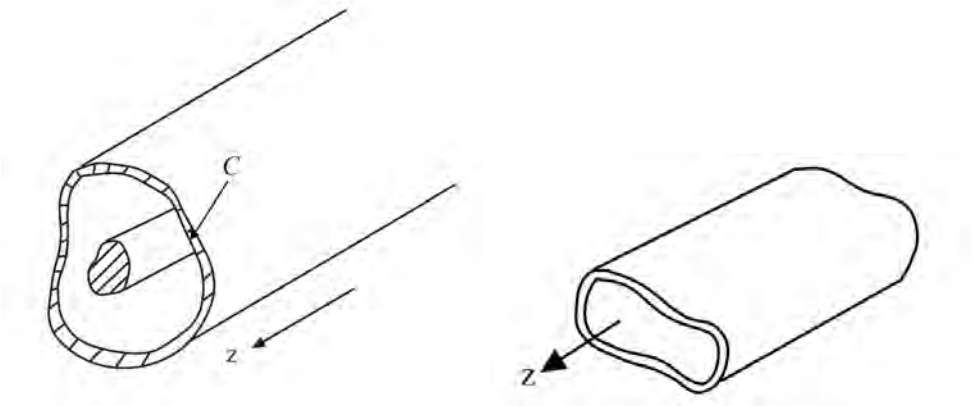


Figure 22.2: A hollow metallic waveguide with a center conductor (left), and without a center conductor (right).

Next, we look at the boundary condition for $\Psi_h(\mathbf{r})$ which is derivable from the boundary condition for $\mathbf{E}$. The boundary condition for $\mathbf{E}$ is that $\hat{n} \times \mathbf{E} = 0$ on C, the PEC wall of the waveguide. But from (22.1.3), using the BAC-CAB formula, then

$$\hat{n} \times \mathbf{E} = \hat{n} \times (\nabla \times \hat{z}\Psi_h) = -\hat{n} \cdot \nabla \Psi_h = 0 \tag{22.1.11}$$

(In applying the BAC-CAB formula, one has to be mindful that ∇ operates on a function to its right, and the function Ψ_h should be placed to the right of the ∇ operator always.)

In the above $\hat{n} \cdot \nabla = \hat{n} \cdot \nabla_s$ where $\nabla_s = \hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y}$ (a 2D gradient operator) since $\hat{n}$ has no z component. The boundary condition (22.1.11) then becomes

$$\hat{n} \cdot \nabla_s \Psi_h = \partial_n \Psi_h = 0, \text{ on } C \tag{22.1.12}$$

where C is the waveguide wall where ∂_n is a shorthand notation for $\hat{n} \cdot \nabla_s$ operator which is a scalar operator for normal derivative. The above is also known as the homogeneous Neumann boundary condition.

Furthermore, in a waveguide, just as in a transmission line case, we are looking for traveling wave solutions of the form $\exp(\mp j\beta_z z)$ for (22.1.10), or that

$$\Psi_h(\mathbf{r}) = \Psi_{hs}(\mathbf{r}_s)e^{\mp j\beta_z z} \tag{22.1.13}$$

where $\mathbf{r}_s = \hat{x}x + \hat{y}y$,[6] or in short, $\Psi_{hs}(\mathbf{r}_s) = \Psi_{hs}(x, y)$ is a 2D function. Thus from the above, $\partial_n \Psi_h = 0$ implies that $\partial_n \Psi_{hs} = 0$ since ∂_n involves ∂_x and ∂_y, and only $\Psi_{hs}(x, y)$ is a function of

[5] (22.1.10) is a sufficient but not necessary condition.

[6] In waveguide theory, we will often use "s" to denote a quantity that is transverse to the z axis, the direction of propagation of the traveling wave.

x and y. With this assumption, $\frac{\partial^2}{\partial z^2} \to -{\beta_z}^2$, and (22.1.10) becomes even simpler, namely that,

$$({\nabla_s}^2 + \beta^2 - {\beta_z}^2)\Psi_{hs}(\mathbf{r}_s) = ({\nabla_s}^2 + \beta_s^2)\Psi_{hs}(\mathbf{r}_s) = 0 \ , \quad \partial_n \Psi_{hs}(\mathbf{r}_s) = 0, \text{ on } C \tag{22.1.14}$$

where $\nabla_s^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ and $\beta_s^2 = \beta^2 - \beta_z^2$. The above is a boundary value problem (BVP) for a 2D waveguide problem. The above 2D wave equation (or partial differential equation) is also called the reduced wave equation.

Please notice that the above equation, rewritten as

$$-{\nabla_s}^2 \Psi_{hs}(\mathbf{r}_s) = +\beta_s^2 \Psi_{hs}(\mathbf{r}_s) \tag{22.1.15}$$

is "homomorphic" to the matrix eigenvalue problem $\overline{\mathbf{A}} \cdot \mathbf{x} = \lambda \mathbf{x}$ where ∇_s^2 is playing the role of the matrix operator, while $\Psi_{hs}(\mathbf{r}_s)$ is analogous to the eigenvector $\mathbf{x}$ and β_s^2 is similar to the eigenvalue λ. The difference from the matrix operator is that the differential operator ∇_s^2 does not have a unique inverse unless boundary conditions are specified.

22.1.3 TM Case ($E_z \neq 0$, $H_z = 0$, TM$_z$ Case)

Repeating similar treatment for TM waves, the TM magnetic field is then

$$\mathbf{H} = \nabla \times \hat{z} \Psi_e(\mathbf{r}) \tag{22.1.16}$$

where

$$(\nabla^2 + \beta^2)\Psi_e(\mathbf{r}) = 0 \tag{22.1.17}$$

The subscript e is used for the pilot potential because Ψ_e can be related to the z component of the $\mathbf{E}$ field. We need to derive the boundary condition for $\Psi_e(\mathbf{r})$ from the fundamental boundary condition that $\hat{n} \times \mathbf{E} = 0$ on the PEC waveguide wall. To this end, we find the corresponding $\mathbf{E}$ field by taking the curl of the magnetic field in (22.1.16), and thus the $\mathbf{E}$ field is proportional to

$$\mathbf{E} \sim \nabla \times \nabla \times \hat{z} \Psi_e(\mathbf{r}) = \nabla \nabla \cdot (\hat{z} \Psi_e) - \nabla^2 \hat{z} \Psi_e = \nabla \frac{\partial}{\partial z} \Psi_e + \hat{z} \beta^2 \Psi_e \tag{22.1.18}$$

where we have used the BAC-CAB formula to simplify the above. The tangential component of the above is $\hat{n} \times \mathbf{E}$ which is proportional to

$$\hat{n} \times \nabla \frac{\partial}{\partial z} \Psi_e + \hat{n} \times \hat{z} \beta^2 \Psi_e$$

In the above, $\hat{n} \times \nabla$ is a tangential derivative, and it is clear that both the above terms will be zero if $\Psi_e = 0$ on the waveguide wall. Therefore, if

$$\Psi_e(\mathbf{r}) = 0 \text{ on } C, \tag{22.1.19}$$

where C is the waveguide wall, then,

$$\hat{n} \times \mathbf{E}(\mathbf{r}) = 0 \text{ on } C \tag{22.1.20}$$

Equation (22.1.19) is also called the homogeneous Dirichlet boundary condition.

Next, we assume that

$$\Psi_e(\mathbf{r}) = \Psi_{es}(\mathbf{r}_s)e^{\mp j\beta_z z} \tag{22.1.21}$$

This will allow us to replace $\partial^2/\partial z^2 = -\beta_z^2$ in (22.1.17). With some manipulation, the boundary value problem (BVP) related to equation (22.1.17) reduces to a simpler partial differential equation for a 2D problem, viz., a reduced wave equation,

$$(\nabla_s{}^2 + \beta_s^2)\Psi_{es}(\mathbf{r}_s) = 0 \tag{22.1.22}$$

with the homogeneous Dirichlet boundary condition that

$$\Psi_{es}(\mathbf{r}_s) = 0, \, \mathbf{r}_s \text{ on } C \tag{22.1.23}$$

The above summarizes the basic theory for hollow waveguides of arbitrary shape. To illustrate the above theory, we will solve some simple waveguide problems.

22.2 Rectangular Waveguides

Rectangular waveguides are among the simplest waveguides to analyze because closed form solutions exist in cartesian coordinates, the simplest of coordinate systems. One can imagine traveling waves in the x and y directions bouncing off the walls of the waveguide causing standing waves to exist inside the waveguide. We have already seen this wave physics in a transmission line: when a transmission line is terminated with a short, traveling waves in both z directions are observed, giving rise to a standing wave. But in a rectangular waveguide, we will see standing waves in the x and y directions, and a traveling wave in the z direction.

As shall be shown, it turns out that not all electromagnetic waves can be guided by a hollow waveguide. Only when the wavelength is short enough, or the frequency is high enough that an electromagnetic wave can be guided by a hollow waveguide. When a waveguide mode cannot propagate in a waveguide, that mode is regarded as cut-off. The concept of cut-off for hollow waveguide is quite different from that of a dielectric waveguide we have studied previously.

22.2.1 TE Modes ($H_z \neq 0$, H Modes or TE$_z$ Modes)

For this mode, the scalar potential $\Psi_{hs}(\mathbf{r}_s)$ satisfies

$$(\nabla_s{}^2 + \beta_s{}^2)\Psi_{hs}(\mathbf{r}_s) = 0, \quad \frac{\partial}{\partial n}\Psi_{hs}(\mathbf{r}_s) = 0 \quad \text{on } C \tag{22.2.1}$$

where $\beta_s{}^2 = \beta^2 - \beta_z{}^2$. A viable solution using separation of variables[7] for $\Psi_{hs}(x, y)$ is then

$$\Psi_{hs}(x, y) = A\cos(\beta_x x)\cos(\beta_y y) \tag{22.2.2}$$

where $\beta_x{}^2 + \beta_y^2 = \beta_s^2$. One can see that the above is the representation of standing waves in the x and y directions. It is quite clear that $\Psi_{hs}(x, y)$ satisfies the BVP (boundary value problem)

[7]For those who are not familiar with this topic, please consult p. 385 of Kong [34].

and boundary conditions defined by equation (22.2.1). Furthermore, cosine functions, rather than sine functions are chosen with the hindsight that the above satisfies the homogenous Neumann boundary condition at $x = 0$, and $y = 0$ surfaces.

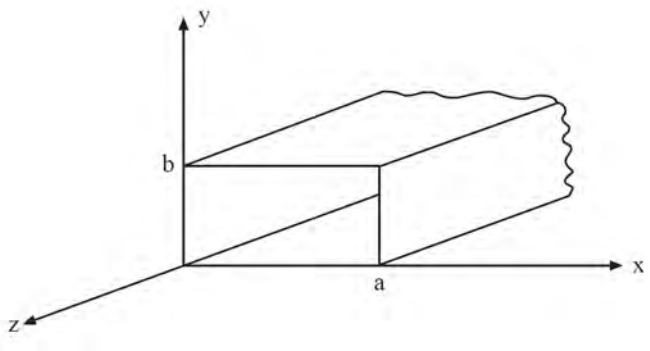


Figure 22.3: The schematic of a rectangular waveguide. By convention, the length of the longer side is usually named a.

To further satisfy the boundary condition at $x = a$, and $y = b$ surfaces, it is necessary that the boundary condition for eq. (22.1.12) is satisfied or that

$$\partial_x \Psi_{hs}(x, y)|_{x=a} \sim \sin(\beta_x a) \cos(\beta_y y) = 0, \tag{22.2.3}$$

$$\partial_y \Psi_{hs}(x, y)|_{y=b} \sim \cos(\beta_x x) \sin(\beta_y b) = 0, \tag{22.2.4}$$

The above puts constraints on the values of β_x and β_y, implying that $\beta_x a = m\pi$, $\beta_y b = n\pi$ where m and n are integers. Hence, (22.2.2) becomes

$$\Psi_{hs}(x, y) = A \cos\left(\frac{m\pi}{a} x\right) \cos\left(\frac{n\pi}{b} y\right) \tag{22.2.5}$$

where $\beta_x = \frac{m\pi}{a}$, $\beta_y = \frac{n\pi}{b}$. They can only take on these values in order for the boundary conditions to be satisfied. Consequently,

$$\beta_x^2 + \beta_y^2 = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 = \beta_s^2 = \beta^2 - {\beta_z}^2 \tag{22.2.6}$$

Clearly, (22.2.5) satisfies the requisite homogeneous Neumann boundary condition at the four waveguide walls.

At this point, it is prudent to stop and ponder on what we have done. Equation (22.2.1) is homomorphic to a matrix eigenvalue problem

$$\overline{\mathbf{A}} \cdot \mathbf{x}_i = \lambda_i \mathbf{x}_i \tag{22.2.7}$$

where $\mathbf{x}_i$ is the eigenvector and λ_i is the eigenvalue. Therefore, β_s^2 is actually an eigenvalue, and $\Psi_{hs}(\mathbf{r}_s)$ is an eigenfunction (or an eigenmode), which is analogous to an eigenvector. Here, the

eigenvalue β_s^2 is indexed by m, n, so is the eigenfunction in (22.2.5): it is also indexed by m and n. The corresponding eigenmode is also called the TE_{mn} mode.

The above condition on β_s^2 expressed by (22.2.6) is also known as the guidance condition for the modes in the waveguide. Furthermore, from (22.2.6),

$$\beta_z = \sqrt{\beta^2 - \beta_s^2} = \sqrt{\beta^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \tag{22.2.8}$$

And from (22.2.8), when the frequency is low enough, then

$$\beta^2 = \omega^2\mu\varepsilon < \beta_s^2 = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 \tag{22.2.9}$$

and β_z in (22.2.8) becomes pure imaginary and the mode cannot propagate or becomes evanescent in the z direction.[8] For fixed m and n, the frequency at which the above happens is called the cut-off frequency of the TE_{mn} mode of the waveguide. It is given by

$$\omega_{mn,c} = \frac{1}{\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2} \tag{22.2.10}$$

When $\omega < \omega_{mn,c}$, or the wavelength is longer than a certain critical value,[9] the TE_{mn} mode is evanescent and cannot propagate inside the waveguide. The corresponding cut-off wavelength is then

$$\lambda_{mn,c} = \frac{2}{[\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2]^{1/2}} \tag{22.2.11}$$

So when $\lambda > \lambda_{mn,c}$, the mode cannot propagate inside the waveguide or it cannot "enter" the waveguide.

Lowest Guided TE Mode in a Rectangular Waveguide

When $m = n = 0$, then $\Psi_h(\mathbf{r}) = \Psi_{hs}(x, y)\exp(\mp j\beta_z z)$ from (22.2.5) is a function independent of x and y. Then $\mathbf{E}(\mathbf{r}) = \nabla \times \hat{z}\Psi_h(\mathbf{r}) = \nabla_s \times \hat{z}\Psi_h(\mathbf{r}) = 0$. It turns out the only way for $H_z \neq 0$ is for $\mathbf{H}(\mathbf{r}) = \hat{z}H_0$ which is a static field in the waveguide. This is not a very interesting mode, and thus TE_{00} propagating mode is assumed not to exist and not useful. So the TE_{mn} modes cannot have both $m = n = 0$.

As such, the TE_{10} mode, when $a > b$, is the mode with the lowest cut-off frequency or longest cut-off wavelength. Only when the frequency is above this cut-off frequency and the wavelength is shorter than this cut-off wavelength, can then be the TE_{10} mode propagate.

For the TE_{10} mode, for the mode to propagate, from (22.2.11), it is needed that

$$\lambda < \lambda_{10,c} = 2a \tag{22.2.12}$$

[8] We have seen this happening in a plasma medium earlier and also in total internal reflection.

[9] We use the formula that $\beta = \omega/v = 2\pi/\lambda$ to find the wavelength, assuming a plane wave propagating in the same homogeneous medium that fills the waveguide.

The above has the nice physical meaning that the wavelength has to be smaller than $2a$ in order for the mode to fit into the waveguide. As a mnemonic, we can think that photons have "sizes", corresponding to its wavelength. Only when its wavelength is small enough can the photons go into (or be guided by) the waveguide. The TE_{10} mode, when $a > b$, is also the mode with the lowest cut-off frequency or longest cut-off wavelength.

It is seen with the above analysis, when the wavelength is short enough, or frequency is high enough, many modes can be guided. Each of these modes has a different group and phase velocity. But for most applications, only a single guided mode is desirable. Hence, the knowledge of the cut-off frequencies of the fundamental mode (the mode with the lowest cut-off frequency) and the next higher mode is important. This allows one to pick a frequency window within which only a single mode can propagate in the waveguide.

It is to be noted that when a mode is cut-off, the field is evanescent, and there is no real power flow down the waveguide: Only reactive power is carried by such a mode as the evanescent wave in total internal reflection (see exercise problem in Chapter 18).

Exercises for Lecture 22

Problem 22-1: Find the group and phase velocity of a guided mode in a hollow waveguide.

(i) Show that their product is always a constant.

(ii) Explain why the group velocities are zero and the phase velocities are infinite right at cut-off of a mode.

(iii) What do the group and phase velocities become when the frequency is very high? Explain why? (Hint: Visualize the orientation of the $\boldsymbol{\beta}$ vector in a hollow waveguide as the frequency becomes very large.)

Chapter 23

More on Hollow Waveguides

We have seen that the hollow waveguide is one of the simplest of waveguides other than the transmission line. Closed form solutions exist for such waveguides as seen in the rectangular waveguide case. The solution is elegantly simple and beautiful requiring only trigonometric functions. So we will continue with the study of the rectangular waveguide, and then address another waveguide, the circular waveguide where closed form solutions also exist. However, the solution has to be expressed in terms of "Bessel functions", called special functions. As the name implies, these functions are seldom used outside the context of studying wave phenomena. Bessel functions in cylindrical coordinates are the close cousin of the sinusoidal functions in cartesian coordinates. Whether Bessel functions are more complex or esoteric compared to sinusoidal functions is in the eyes of the beholder. Once one becomes familiar with them, they are simple! They are also the function that describes the concentric ripple wave that you see in your tea cup every morning (see Figure 23.1)!

Figure 23.1: The ripple wave (also called capillary wave) in your tea cup is describable by a Bessel function (courtesy of dreamstime.com).

23.1 Rectangular Waveguides, Contd.

We have seen the mathematics for the TE modes of a rectangular waveguide in the previous lecture. We shall now study the TM modes and both the TE and TM modes of a circular waveguide.

23.1.1 TM Modes ($E_z \neq 0$, E Modes or TM_z Modes)

These modes are not the exact dual of the TE modes because of the boundary conditions. The dual of a PEC (perfect electric conducting) wall is a PMC (perfect magnetic conducting) wall. However, the previous exercise for TE modes can be repeated for the TM modes with caution on the boundary conditions. The scalar wave function (or eigenfunction/eigenmode) for the TM modes, satisfying the homogeneous Dirichlet (instead of Neumann)[1] boundary condition with ($\Psi_{es}(\mathbf{r}_s) = 0$) on the entire waveguide wall is

$$\Psi_{es}(x, y) = A \sin\left(\frac{m\pi}{a}x\right) \sin\left(\frac{n\pi}{b}y\right) \tag{23.1.1}$$

In the above, it is to reminded that the meaning of the subscript "e" implies that this scalar potential Ψ is related to the z component of the $\mathbf{E}$ field, while the subscript "s" implies that this refers to a quantity that is transverse to the z axis. Here, $\beta_x = \frac{m\pi}{a}$ and $\beta_y = \frac{n\pi}{b}$. Sine functions are chosen for the standing waves, and the chosen values of β_x and β_y ensure that the Dirichlet boundary condition is satisfied on the $x = a$ and $y = b$ walls. Neither of the m and n can be zero, lest $\Psi_{es}(x, y) = 0$, or the entire field is zero. Hence, both $m > 0$, and $n > 0$ are needed. Thus, the lowest TM mode is the TM_{11} mode. Thinking of this as an eigenvalue problem, as mentioned in the previous lecture, then the eigenvalue is

$$\beta_s^2 = \beta_x^2 + \beta_y^2 = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 \tag{23.1.2}$$

which is the same as the TE case. Therefore, the corresponding cut-off frequencies and cut-off wavelengths for the TM_{mn} modes are the same as the TE_{mn} modes. Also, these TE and TM modes are degenerate when they share the same eigevalues. In other words, the lowest mode, which is the TM_{11} mode and the TE_{11} have the same cut-off frequency. Figure 23.2 shows the dispersion curves for different modes of a rectangular waveguide. Notice that the group velocities of all the modes are zero at cut-off, and then the group velocities approach that of the waveguide medium inside as frequency becomes large. These observations can be explained physically when we study the bouncing-wave picture next.

[1] Again, "homogeneous" here for the math community means "zero".

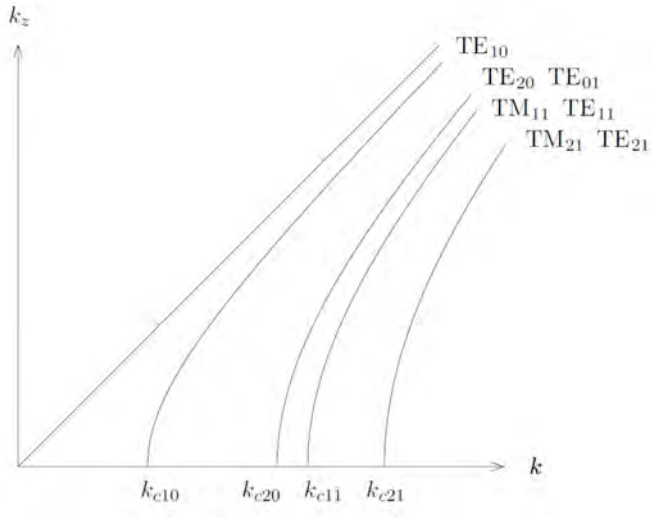


Figure 23.2: Dispersion curves for a rectangular waveguide (courtesy of J.A. Kong [34]). Notice that the lowest TM mode is the TM_{11} mode, and that the TM_{mn} modes and the TE_{mn} modes have the same cut-off frequencies if they exist. Here, k is equivalent to β in this course. At cut-off, $\beta_z = k_z = 0$, or the guided mode does not propagate in the z direction, and as shown above, the group velocity is zero. But when $\omega \to \infty$, the mode propagates in direction almost parallel to the axis of the waveguide (this is termed paraxial wave in the parlance of wave physics), and hence, the group velocity approaches that of the waveguide medium.

23.1.2 Bouncing Wave Picture

We have seen that the transverse variation of a mode in a rectangular waveguide can be expanded in terms of sine and cosine functions which represent standing waves which are superposition of two traveling waves, or they are

$$\begin{aligned}\left[e^{-j\beta_x x} \pm e^{j\beta_x x}\right]\left[e^{-j\beta_y y} \pm e^{j\beta_y y}\right] \\ = e^{-j\beta_x x - j\beta_y y} + e^{j\beta_x x + j\beta_y y} \pm e^{j\beta_x x - j\beta_y y} \pm e^{-j\beta_x x + j\beta_y y}\end{aligned} \tag{23.1.3}$$

Each term on the right-hand side corresponds to a plane wave travelling in different directions. When the above is expanded and together with the $\exp(-j\beta_z z)$, the mode is propagating in the z direction in addition to being the standing waves in the transverse direction. Or we see four waves bouncing around in the x and y directions and propagating in the z direction. The picture of this bouncing wave is depicted in Figure 23.3.

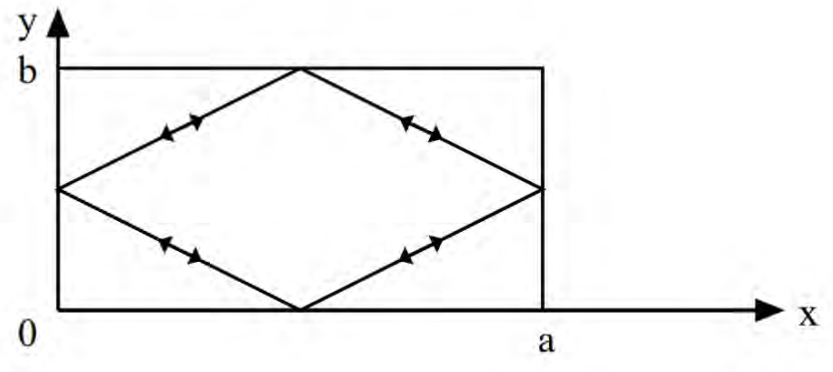


Figure 23.3: The waves in a rectangular waveguide can be thought of as bouncing waves off the four walls as they propagate in the z direction out of the paper.

23.1.3 Field Plots

Given the knowledge of the vector pilot potential of a waveguide, one can derive all the field components. For example, for the TE modes, if we know $\Psi_h(\mathbf{r})$, then

$$\mathbf{E} = \nabla \times \hat{z}\Psi_h(\mathbf{r}), \qquad \mathbf{H} = -\nabla \times \mathbf{E}/(j\omega\mu) \tag{23.1.4}$$

Then all the electromagnetic field of a waveguide mode can be found and computed for these modes, and similarly for TM modes.

Plots of the fields of different rectangular waveguide modes are shown in Figure 23.4. Notice that for higher m's and n's, with $\beta_x = m\pi/a$ and $\beta_y = n\pi/b$, the corresponding β_x and β_y are larger with higher spatial frequencies. Thus, the transverse spatial wavelengths are getting shorter. Also, since $\beta_z = \sqrt{\beta^2 - \beta_x^2 - \beta_y^2} = \sqrt{\beta^2 - (m\pi/a)^2 - (n\pi/b)^2}$, higher frequencies where $\beta^2 = \omega^2\mu\varepsilon$ are larger than $(m\pi/a)^2 + (n\pi/b)^2$ to make β_z real in order to propagate the higher order modes or the high m and n modes in a rectangular waveguide.

Notice also how the electric field and magnetic field curl around each other. Since $\nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E}$ and $\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H}$, they do not curl around each other "immediately" in the time domain, but with a $\pi/2$ phase delay due to the $j\omega$ factor corresponding to a time delay in the time domain. Therefore, in a snapshot of the $\mathbf{E}$ and $\mathbf{H}$ fields shown in Figure 23.5: they do not curl around each other at one location, but at a displaced location due to the $\pi/2$ phase difference due to $j\omega$ giving rise to a time delay.

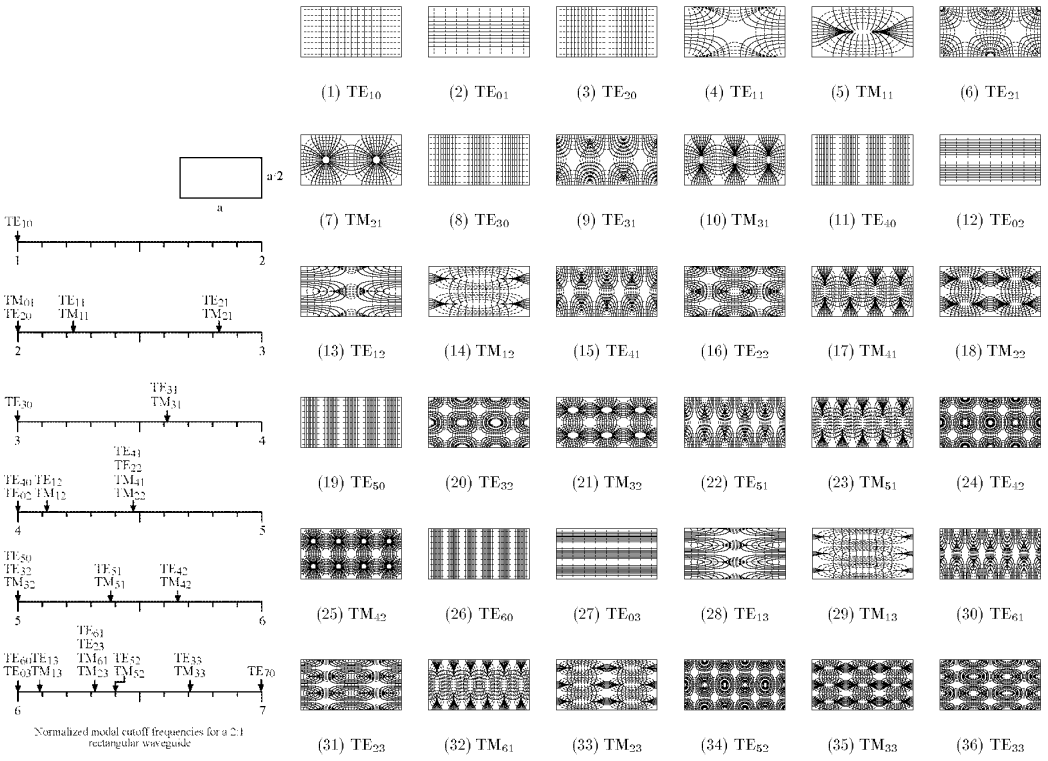


Figure 23.4: Transverse field plots of different modes in a rectangular waveguide (courtesy of Andy Greenwood. Original plots published in Lee, Lee, and Chuang, IEEE T-MTT, 33.3 (1985): pp. 271-274. [161]).

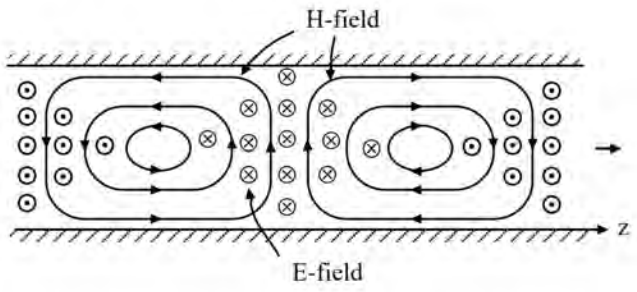


Figure 23.5: A snapshot of the field plot of a TE_{10} mode propagating in the z direction of a rectangular waveguide. Notice that the **E** and **H** fields do not exactly curl around each other because of the time delay.

23.2 Circular Waveguides

Another waveguide whose closed-form solutions can be easily obtained is the circular hollow waveguide as shown in Figure 23.6. Now they involve the use of Bessel functions which are special functions. They are different from the trigonometric functions which are used pervasively.

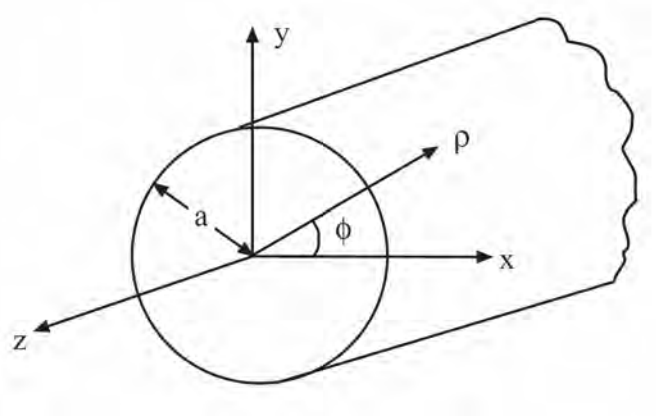


Figure 23.6: Schematic of a circular waveguide in cylindrical coordinates. It is one of the separable coordinate systems.

23.2.1 TE Case

For a circular waveguide, it is best first to express the Laplacian operator, ${\nabla_s}^2 = \nabla_s \cdot \nabla_s$, in cylindrical coordinates. The second term ∇_s is a gradient operator while the first term $\nabla_s \cdot$ is a divergence operator: they have different physical meanings. Formulas for grad and div operators in cylindrical coordinates are given in many text books [34, 162]. Doing a table lookup,

$$\nabla_s \Psi = \hat{\rho} \frac{\partial}{\partial \rho} \Psi + \hat{\phi} \frac{1}{\rho} \frac{\partial}{\partial \phi}$$

$$\nabla_s \cdot \mathbf{A} = \frac{1}{\rho} \frac{\partial}{\partial \rho} \rho A_\rho + \frac{1}{\rho} \frac{\partial}{\partial \phi} A_\phi$$

Then

$$\left({\nabla_s}^2 + {\beta_s}^2\right) \Psi_{hs} = \left(\frac{1}{\rho} \frac{\partial}{\partial \rho} \rho \frac{\partial}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2}{\partial \phi^2} + {\beta_s}^2 \right) \Psi_{hs}(\rho, \phi) = 0 \tag{23.2.1}$$

The above is the partial differential equation for field in a circular waveguide. It is an eigenvalue problem where β_s^2 is the eigenvalue, and $\Psi_{hs}(\mathbf{r}_s)$ is the eigenfunction (equivalence of an eigenvector). Here, $\mathbf{r}_s = \hat{\rho}\rho + \hat{\phi}\phi$ is the transverse to z position vector in cylindrical coordinates in 2D. Please be reminded that the above is "homomorphic" to a matrix eigenvalue problem as discussed at the end of Section 22.1.2 in the previous lecture.

Using separation of variables, we let

$$\Psi_{hs}(\rho,\phi) = B_n(\beta_s\rho)e^{\pm jn\phi} \tag{23.2.2}$$

Then $\frac{\partial^2}{\partial\phi^2} \to -n^2$, and (23.2.1) simplifies to an ordinary differential equation which is

$$\left(\frac{1}{\rho}\frac{d}{d\rho}\rho\frac{d}{d\rho} - \frac{n^2}{\rho^2} + {\beta_s}^2\right) B_n(\beta_s\rho) = 0 \tag{23.2.3}$$

Here, dividing the above equation by β_s^2, we can let $\beta_s\rho$ in (23.2.2) and (23.2.3) be x. Then the above can be rewritten as

$$\left(\frac{1}{x}\frac{d}{dx}x\frac{d}{dx} - \frac{n^2}{x^2} + 1\right) B_n(x) = 0 \tag{23.2.4}$$

The above is known as the Bessel equation whose solutions are special functions denoted as $B_n(x)$.[2]

These special functions, also called cylinder functions, are $J_n(x)$, $N_n(x)$, $H_n^{(1)}(x)$, and $H_n^{(2)}(x)$ which are called Bessel, Neumann, Hankel function of the first kind, and Hankel function of the second kind, respectively, where n is their order, and x is their arguments.[3] Since this is a second order ordinary differential equation, it has only two independent solutions. Therefore, two of the four commonly encountered solutions of Bessel equation are independent. Thus, they can be expressed in terms of each other. Their relationships are shown below:[4]

$$\text{Bessel,} \qquad J_n(x) = \frac{1}{2}[{H_n}^{(1)}(x) + {H_n}^{(2)}(x)] \tag{23.2.5}$$

$$\text{Neumann,} \qquad N_n(x) = \frac{1}{2j}[{H_n}^{(1)}(x) - {H_n}^{(2)}(x)] \tag{23.2.6}$$

$$\text{Hankel–First Kind,} \qquad {H_n}^{(1)}(x) = J_n(x) + jN_n(x) \tag{23.2.7}$$

$$\text{Hankel–Second Kind,} \qquad {H_n}^{(2)}(x) = J_n(x) - jN_n(x) \tag{23.2.8}$$

When x is large, using the asymptotic formulas for these special functions [113], it can be shown that

$${H_n}^{(1)}(x) \sim \sqrt{\frac{2}{\pi x}}e^{jx-j(n+\frac{1}{2})\frac{\pi}{2}}, \quad x \to \infty \tag{23.2.9}$$

$${H_n}^{(2)}(x) \sim \sqrt{\frac{2}{\pi x}}e^{-jx+j(n+\frac{1}{2})\frac{\pi}{2}}, \quad x \to \infty \tag{23.2.10}$$

They correspond to traveling wave solutions when $x = \beta_s\rho \to \infty$.

Since $J_n(x)$ and $N_n(x)$ are linear superpositions of these traveling wave solutions, they correspond to standing wave solutions. Moreover, $N_n(x)$, ${H_n}^{(1)}(x)$, and ${H_n}^{(2)}(x)$ are singular at $x = 0$. Or in a word, they tend to ∞ when $x \to 0$. Since the field has to be regular when $\rho \to 0$ at the

[2]Studied by Friedrich Wilhelm Bessel, 1784-1846.

[3]Some textbooks use $Y_n(x)$ for Neumann functions.

[4]Their relations with each other are similar to those between $\exp(\pm jx)$, $\sin(x)$, and $\cos(x)$. The marked difference is that the trigonometric functions are all regular, but some of the cylinder special functions are singular at $x = 0$.

center of the waveguide shown in Figure 23.6, the only viable solution for the hollow waveguide, to be chosen from (23.2.5) to (23.2.9), is that $B_n(\beta_s\rho) = AJ_n(\beta_s\rho)$ which is regular at $\rho = 0$. Thus for a circular hollow waveguide, the eigenfunction or mode has to be of the form

$$\Psi_{hs}(\rho, \phi) = AJ_n(\beta_s\rho)e^{\pm jn\phi} \tag{23.2.11}$$

To ensure that the eigenfunction and the eigenvalue are unique, boundary condition for the partial differential equation is needed. The homogeneous Neumann boundary condition,[5] or that $\partial_n\Psi_{hs} = 0$, on the PEC waveguide wall then translates to

$$\frac{d}{d\rho}J_n(\beta_s\rho) = 0, \quad \rho = a \tag{23.2.12}$$

Defining ${J_n}'(x) = \frac{d}{dx}J_n(x)$,[6] the above is the same as

$${J_n}'(\beta_s a) = 0 \tag{23.2.13}$$

The above are the zeros of the derivative of Bessel functions and they are tabulated in many textbooks and handbooks.[7] The m-th zero of ${J_n}'(x)$ is denoted here to be ξ_{nm}, a dimensionless number. Plots of Bessel functions and their derivatives are shown in Figure 23.8, and some zeros of Bessel functions and their derivatives are also shown in Tables 23.2.1 and 23.2.2. With this knowledge, the guidance condition for a waveguide mode is then

$$\beta_s = \xi_{nm}/a \tag{23.2.14}$$

for the TE_{nm} mode. From the above, β_s^2 can be obtained which is the eigenvalue of (23.2.1) and (23.2.3). It is a constant independent of frequency.

Using the fact that $\beta_z = \sqrt{\beta^2 - \beta_s^2}$, then β_z will become pure imaginary if $\beta = \omega\sqrt{\mu\varepsilon}$ is small. Here, β can be made small by lowering the frequency low so that $\beta^2 < \beta_s^2$. From this, the corresponding cut-off frequency of the TE_{nm} mode is

$$\omega_{nm,c} = \frac{1}{\sqrt{\mu\varepsilon}}\frac{\xi_{nm}}{a} \tag{23.2.15}$$

When $\omega < \omega_{nm,c}$, the corresponding mode cannot propagate in the waveguide as β_z becomes pure imaginary. The corresponding cut-off wavelength, using $\omega\sqrt{\mu\varepsilon} = \beta = 2\pi/\lambda$, is then

$$\lambda_{nm,c} = \frac{2\pi}{\xi_{nm}}a \tag{23.2.16}$$

By the same token, when $\lambda > \lambda_{nm,c}$, the corresponding mode cannot be guided by the waveguide. It is not exactly precise to say this, but this gives us the heuristic notion that if wavelength or "size" of the wave or photon is too big, it cannot fit inside the waveguide.

[5] Note that "homogeneous" here means "zero" in math.

[6] Note that this is a standard math notation, which has a different meaning in some engineering texts.

[7] Notably, Abramowitz and Stegun, Handbook of Mathematical Functions [113]. An online version is available at [163].

23.2.2 TM Case

The corresponding partial differential equation and boundary value problem for this case is[8]

$$\left(\frac{1}{\rho}\frac{\partial}{\partial\rho}\rho\frac{\partial}{\partial\rho}+\frac{1}{\rho^2}\frac{\partial^2}{\partial\phi^2}+{\beta_s}^2\right)\Psi_{es}(\rho,\phi)=0 \tag{23.2.17}$$

with the homogeneous Dirichlet boundary condition, $\Psi_{es}(a,\phi)=0$, on the waveguide wall.[9] The eigenfunction solution is

$$\Psi_{es}(\rho,\phi)=AJ_n(\beta_s\rho)e^{\pm jn\phi} \tag{23.2.18}$$

with the boundary condition that $J_n(\beta_s a)=0$. The m-th zeros of $J_n(x)$ are labeled as α_{nm} here [113, 34], as well as in Tables 23.2.1 and 23.2.2; and hence, the guidance condition for the TM_{nm} mode is that

$$\beta_s=\frac{\alpha_{nm}}{a} \tag{23.2.19}$$

where the eigenvalue for (23.2.17) is β_s^2 which is a constant independent of frequency. With $\beta_z=\sqrt{\beta^2-\beta_s^2}$, the corresponding cut-off frequency is

$$\omega_{nm,c}=\frac{1}{\sqrt{\mu\varepsilon}}\frac{\alpha_{nm}}{a} \tag{23.2.20}$$

or when $\omega<\omega_{nm,c}$, the mode cannot be guided. The cut-off wavelength is then

$$\lambda_{nm,c}=\frac{2\pi}{\alpha_{nm}}a \tag{23.2.21}$$

with the notion that when $\lambda>\lambda_{nm,c}$, the mode cannot be guided.

Tables 23.2.1 and 23.2.2 show the zeros ξ_{nm} and α_{nm}. They are important for figuring out the cut-off frequencies of the TE and TM modes of a circular hollow waveguide.

It turns out that the lowest mode in a circular waveguide is the TE_{11} mode. It is actually a close cousin of the TE_{10} mode of a rectangular waveguide. This can be gathered by comparing their field plots: these modes morph into each other as we deform the shape of a rectangular waveguide into a circular waveguide.

[8] Again, this is analogous to the matrix eigenvalue problem where differential operators are matrix operators, β_s^2 is the eigenvalue, and the eigenfunction Ψ_{es} is the eigenvector.

[9] Again, note that this problem is analogous to the matrix eigenvalue problem. The matrix has a unique inverse (or is full rank) only if boundary condition is stipulated for the differential operator.

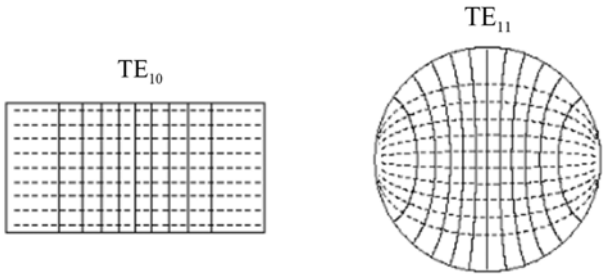


Figure 23.7: Side-by-side comparison of the field plots of the TE_{10} mode of a rectangular waveguide versus that of the TE_{11} mode of a circular waveguide. If one is imaginative enough, one can see that the field plot of TE_{10} mode morphs into that of TE_{11} mode as we change the waveguide shape. Electric fields are those that have to end on the waveguide walls with $\hat{n} \times \mathbf{E} = 0$.

Table 23.2.1. Roots of $J_n'(x) = 0$.

n	ξ_{n1}	ξ_{n2}	ξ_{n3}	ξ_{n4}
0	3.832	7.016	10.174	13.324
1	1.841	5.331	8.536	11.706
2	3.054	6.706	9.970	13.170
3	4.201	8.015	11.346	14.586
4	5.318	9.282	12.682	15.964
5	6.416	10.520	13.987	17.313

Table 23.2.2. Roots of $J_n(x) = 0$.

n	α_{n1}	α_{n2}	α_{n3}	α_{n4}
0	2.405	5.520	8.654	11.792
1	3.832	7.016	10.174	13.324
2	5.135	8.417	11.620	14.796
3	6.380	9.761	13.015	16.223
4	7.588	11.065	14.373	17.616
5	8.771	12.339	15.700	18.980

Figure 23.8 shows the plots of Bessel function $J_n(x)$ and its derivative $J_n'(x)$. Tables 23.2.1 and 23.2.2 show the roots of $J_n'(x)$ and $J_n(x)$ which are important for determining the cut-off frequencies of the TE and TM modes of circular waveguides. They are useful for determining the guidance conditions of the TE_{nm} mode and TM_{nm} mode of a circular waveguide [90].

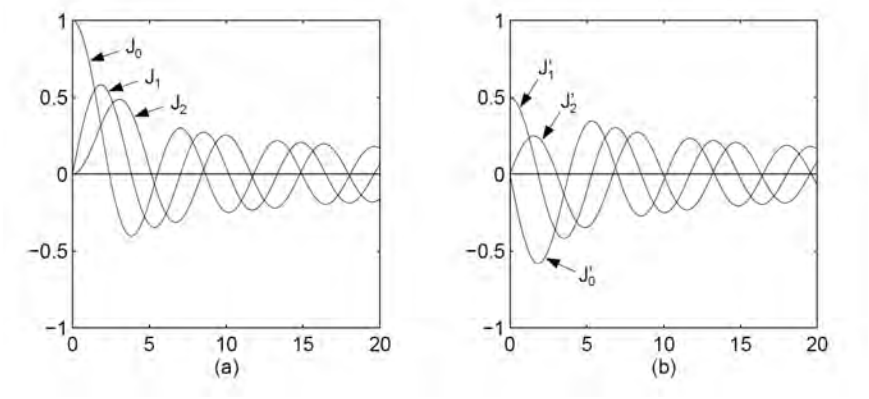


Figure 23.8: Plots of the Bessel functions, $J_n(x)$, of different orders, and their derivatives $J_n'(x)$. The zeros of these functions are used to find the eigenvalue β_s^2 of the problem, and hence, the guidance condition. The left figure is for TM modes, while the right figure is for TE modes. Here, $J_n'(x) = dJ_n(x)/dx$ [90].

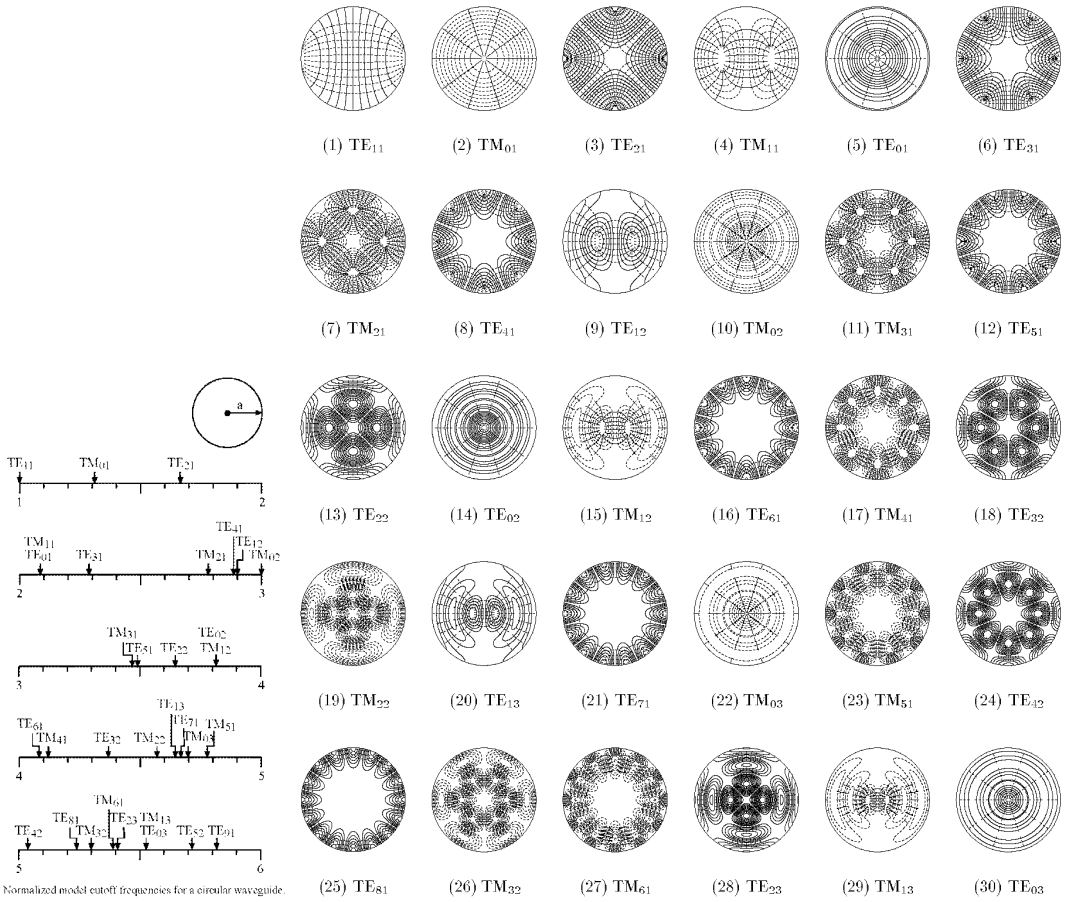


Figure 23.9: Transverse field plots of different modes in a circular waveguide (courtesy of Andy Greenwood. Original plots published in Lee, Lee, and Chuang [161]). The axially symmetric TE_{01} mode has the lowest loss, and finds a number of real-world applications as in radio astronomy.

Exercises for Lecture 23

Problem 23-1:

(i) For a 2 cm by 1 cm rectangular waveguide, and a 1 cm radius circular waveguide, find the first three propagating modes starting with the lowest cutoff frequencies.

(ii) For the TM modes, show that the homogeneous Dirichlet boundary condition that $\Psi_e = 0$ on C, the waveguide wall, all components of the tangential electric field will be zero also.

(iii) For the TM mode, starting with that

$$\mathbf{H}(\mathbf{r}) = \nabla \times \hat{z}\Psi_e(\mathbf{r})$$

Find all the components of electromagnetic fields in a rectangular waveguide.

(iv) Starting with a TM mode in a waveguide where $H_z = 0$, and $E_z \neq 0$ initially, show that if this mode becomes a TEM mode so that $H_z = 0$, $E_z = 0$ then $\beta_z = \beta$. (This can happen in a coaxial waveguide, for instance.) What happens to β_s for the TEM mode, and what happens to the Helmholtz equations that $\Psi_e(\mathbf{r})$ and $\Psi_h(\mathbf{r})$ originally satisfy? Explain why the fields of the TEM mode of a waveguide, if it exists, is electrostatic and magnetostatic in nature.

Chapter 24

More on Waveguides and Transmission Lines

Waveguide is a fundamental component of microwave circuits and systems. The study of closed form solutions offers us physical insight. One can use such insight to design more complex engineering systems. In this chapter, we will use heuristics to understand some complex systems whose designs follow from physical insight of simpler systems.

In addition, we will show that the waveguide problem is homomorphic to the transmission line problem. Here again, many transmission line techniques can be used to solve some complex waveguide problems approximately encountered in microwave and optical engineering by adding junction capacitances and inductances from Figure 17.8 in Chapter 17.

24.1 Circular Waveguides, Contd.

The scalar potential (or pilot potential) for the modes in the circular waveguide is expressible as

$$\Psi_{\alpha s}(\rho, \phi) = AJ_n(\beta_s \rho)e^{\pm jn\phi} \tag{24.1.1}$$

where $\alpha = h$ for TE waves and $\alpha = e$ for TM waves.[1] The Bessel function or wave is expressible in terms of Hankel functions as in (23.2.5). Since Hankel functions are inward and outward traveling waves, Bessel functions represent standing waves. Therefore, the Bessel waves can be thought of as bouncing traveling waves as found in the rectangular waveguide case. In the azimuthal direction, one can express $e^{\pm jn\phi}$ as traveling waves in the ϕ direction, or they can be expressed as $\cos(n\phi)$ and $\sin(n\phi)$ which are standing waves in the ϕ direction.

[1] As mentioned before, the pilot potentials are related to the H_z and E_z components of the fields as reflected in the subscripts of the potential. Also, the potential is regular at the center of the waveguide and hence, Bessel function is chosen as the solution.

24.1.1 An Application of Circular Waveguide

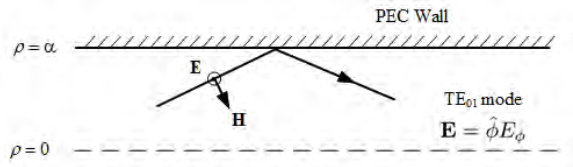


Figure 24.1: Bouncing wave picture of the Bessel wave inside a circular waveguide for the TE_{01} mode. One can also explain the low-loss physics [164] using the TE mode of a parallel-plate waveguide.

When a real-world waveguide is made, the wall of the metal waveguide is not made of perfect electric conductor, but with some metal of finite conductivity. Hence, tangential $\mathbf{E}$ field is not zero on the waveguide wall implying that $\hat{n} \cdot (\mathbf{E} \times \mathbf{H}^*) \neq 0$. Since $\mathbf{E} \times \mathbf{H}^*$ is the complex Poynting's vector, it being not zero implies that energy can dissipate (or power can flow) into the waveguide wall.

It turns out that due to symmetry, the TE_{01} mode of a circular waveguide has the lowest loss of all the waveguide modes counting in even rectangular waveguide modes of a rectangular waveguide. Hence, this waveguide mode is of great interest to astronomers who are interested in building low-loss and low-noise systems.[2]

The TE_{01} mode has electric field given by $\mathbf{E} = \hat{\phi} E_\phi$. Furthermore, looking at the magnetic field, the current is mainly circumferential (azimuthal) flowing in the ϕ direction. By looking at a bouncing wave picture of the guided waveguide mode, this mode, similar to the parallel waveguide mode, has a small component of tangential magnetic field on a waveguide wall: It becomes increasingly smaller as the frequency increases (see Figure 24.1). The reason is that the wave vector for the waveguide becomes increasingly parallel to the axis of the waveguide with a large β_z component compared to the β_s component.[3] In a word, the wave becomes paraxial in the high-frequency limit.

The tangential magnetic field needs to be supported by a surface current on the waveguide wall. This implies that the surface current on the waveguide wall becomes smaller as the frequency increases. Consequently, the wall loss (or copper loss or eddy current loss) of the waveguide becomes smaller for higher frequencies. In fact, for high frequencies, the TE_{01} mode has the smallest copper loss of all waveguide modes (including waveguides of other shapes): It becomes the mode of choice (see Figure 24.2). Waveguides supporting the TE_{01} modes are used to connect the antennas of the very large array (VLA) for detecting extra-terrestrial signals in radio astronomy [168] as shown in Figure 24.3. The low wall loss gives rise to good SNR (signal-to-noise) ratio.[4]

[2]Low-loss systems are also low-noise due to energy conservation and the fluctuation dissipation theorem [159, 160, 165]. This is similar to the Johnson-Nyquist noise. [166, 167]

[3]Recall that for a fixed mode, β_s^2 is the eigenvalue of the system, and hence, β_s is independent of frequency.

[4]This follows from the fluctuation dissipation theorem which is an energy conserving theorem, that says that whatever electromagnetics energy absorbed by the environment is returned to the environment as thermal radiation.

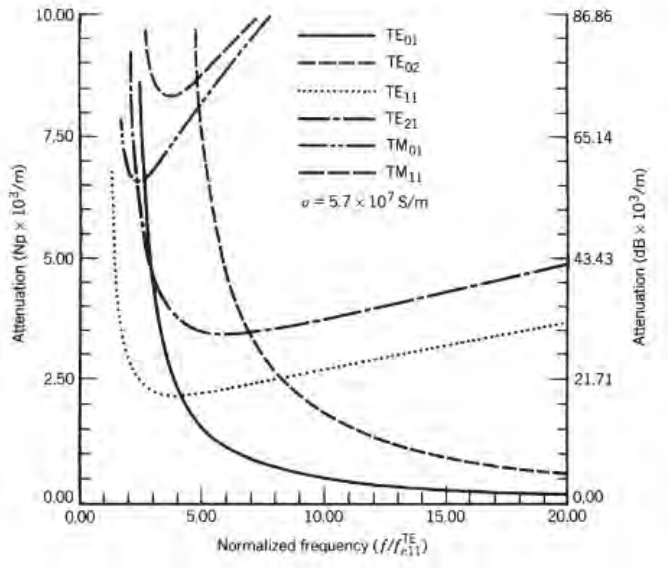


Figure 24.2: Plots of losses of different modes in a circular waveguide of radius 1.5 cm. It is seen that at high frequencies, the TE_{01} mode has the lowest loss (courtesy of [164]).

Figure 24.3: Picture of the Very Large Array in New Mexico, USA (courtesy of [168]). The low loss of the circular waveguide gives good SNR (signal-to-noise ratio) according to the fluctuation dissipation theorem, which makes the system good for detecting weak radio astronomy signals from outer space.

Figure 24.4 shows two ways of engineering a circular waveguide so that the TE_{01} mode is enhanced: (i) by using a mode filter that discourages the guidance of other modes except the TE_{01} mode, and (ii), by designing corrugated waveguide wall to discourage the flow of axial current but encourage the flow of circumferential (azimuthal) current. Thus it discourages the propagation of the non-TE_{01} mode. (The TE_{01} mode only has azimuthal current in the ϕ direction.) More details of circular waveguides can be found in [164]. Typical loss of a circular waveguide can be as low as 2 dB/km.[5]

As shall be shown, an open circular waveguide can be made into an aperture antenna quite easily, because the fields of the aperture are axially symmetric. To this end, the axially symmetric TE_{01} mode is enhanced by design. Because of this, the radiation pattern of such an antenna is axially symmetric, which can be used to produce axially symmetric circularly polarized (CP) waves. Such antenna is called a horn antenna. Ways to enhance the TE_{01} mode of the horn antenna are also desirable [169] as shown in Figure 24.5.

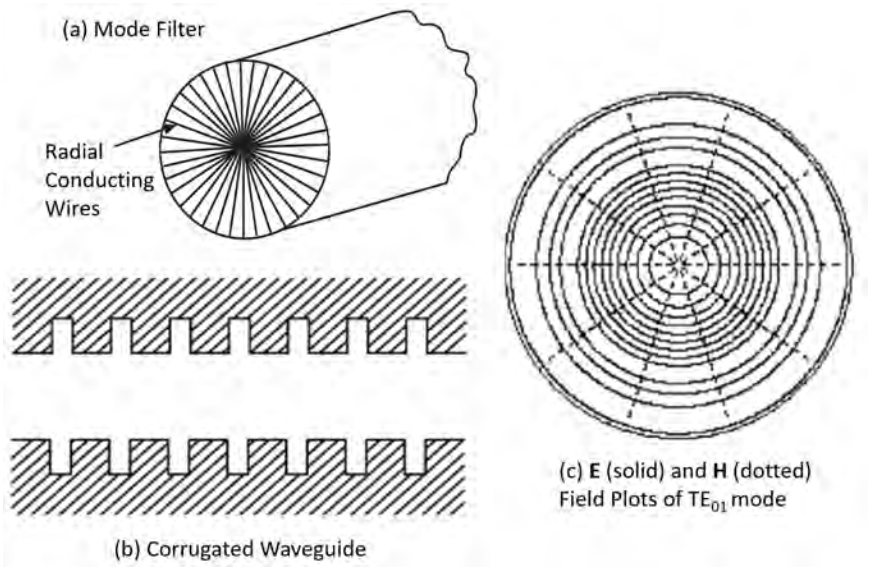


Figure 24.4: The TE_{01} has $\mathbf{E} = \hat{\phi} E_\phi$, and tangential $\mathbf{H}$ field only has a z component on the waveguide wall. Hence, the wall current is purely circumferential (azimuthal) for this mode. There are ways to enhance the TE_{01} mode in a circular waveguide: (a) Using mode filter that only allows the TE_{01} mode to go through. (b) Use corrugated waveguide to discourage axial current flow from the other modes but encourage the circumferential (azimuthal) current flow from this TE_{01} mode. The field plot of the mode is shown in (c). Such waveguide is used in radio astronomy to design the communication links between antennas in a very large array (VLA [168]), or it is used in a circular horn antenna [169].

[5]For optical fiber, this figure of merit (FOM) can be as low as 0.1 dB/km making the optical fiber a darling for long-distance communication.

Figure 24.5: Picture of a circular horn antenna where corrugated wall is used to enhance the TE_{01} mode that is axially symmetric. It discourage the other nonaxially symmetric modes. Therefore, the antenna has an axially symmetric field producing a radiation pattern (to be addressed later) that is axially symmetric (courtesy of [170]).

24.2 Remarks on Quasi-TEM Modes, Hybrid Modes, and Surface Plasmonic Modes

We have analyzed some simple structures where closed form solutions are available. These simple elegant solutions offer physical insight as to how waves are guided, and how they are cut-off from guidance. As has been shown, for some simple waveguides, the modes can be divided into TEM, TE, and TM modes. However, most waveguides are not simple. We will remark on various complexities that arise in real world applications.

24.2.1 Quasi-TEM Modes

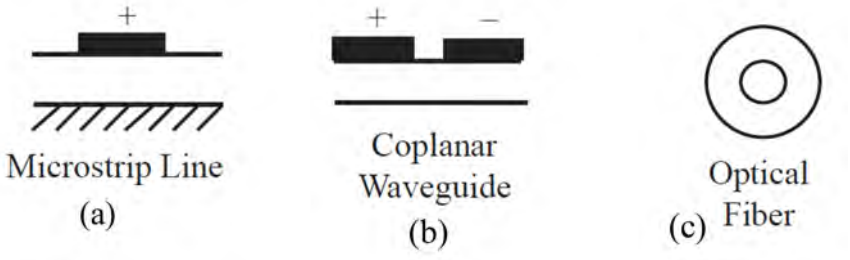


Figure 24.6: Some examples of practical coaxial-like waveguides are microstrip line and co-planar waveguide (left). For the microstrip line (a), the signal line (denoted with a + sign) mimics the center conductor of a coax, while the ground plane (hashed lines) represents the outer conductor of a coax. The coplanar waveguide (b) needs no ground plane, and operates like a twin-ax, where the ± line indicates the signal line and ground line of the twin-ax. The optical fiber is indicated in (c). It operates by total-internal-reflection at the interface between the center (core) of the waveguide, and the cladding (outside the core). The environments of these waveguides are inhomogeneous media, and hence, a pure TEM mode cannot propagate on these waveguides.

Many waveguides cannot support a pure TEM mode even when two conductors are present. For example, two pieces of metal make a transmission line, and in the case of a circular coax, a TEM mode can propagate in the waveguide. But most two-metal transmission lines do not support a pure TEM mode: Instead, they support a quasi-TEM mode. In the optical fiber case, when the index contrast of the fiber is very small, the mode is also quasi-TEM as it has to degenerate to the TEM case when the contrast is absent.

Absence of TEM Modes in Inhomogeneously-Filled Waveguides

In the following, we will give physical arguments as to why a pure TEM mode cannot exists in a microstrip line, a coplanar waveguide, and an optical fiber. When a wave is TEM, it is necessary that the wave propagates with the phase velocity of the medium in which it propagates with $\exp(-j\beta_i z)$ dependence where β_i is the wavenumber of the medium. But when a uniform waveguide has inhomogeneity in between, as shown in Figure 24.6, this is not possible anymore. We can prove this assertion by *reductio ad absurdum*. Very simply put, if the waves are TEM in all the regions, they will have the respective phase velocity of the regions, each with a different $\exp(-j\beta_i z)$ dependence. Thus, phase matching is impossible at the interfaces between these regions.

We shall study this point in greater detail: Assume only TM wave in a piecewise homogeneous region, then using the vector pilot potential approach, the $\mathbf{E}$ field is

$$\mathbf{E} = \frac{1}{j\omega\varepsilon_i}\nabla\times\nabla\times(\hat{z}\Psi_e) \tag{24.2.1}$$

where ε_i is the permittivity of the region. By doing some algebra, and assuming that the field is a waveguide mode such that Ψ_e has $e^{-j\beta_z z}$ dependence, then using the BAC-CAB formula, one can show that the above simplifies to

$$\mathbf{E} = \frac{1}{j\omega\varepsilon_i}\left[\nabla\nabla\cdot\hat{z}\Psi_e - \nabla^2\hat{z}\Psi_e\right] \tag{24.2.2}$$

Or

$$E_z = \frac{1}{j\omega\varepsilon_i}\left[\frac{\partial^2}{\partial z^2}\Psi_e - \nabla^2\Psi_e\right] \tag{24.2.3}$$

Therefore, E_z is given by

$$E_z = \frac{1}{j\omega\varepsilon_i}(\beta_i^2 - \beta_z^2)\Psi_e \tag{24.2.4}$$

The above derivation is certainly valid in a piecewise homogeneous region. But each of the piecewise homogeneous media can be made arbitrary small, and hence, it is also valid for inhomogeneous media. If this mode becomes TEM, then $E_z = 0$ and this is possible only if $\beta_z = \beta_i$. In other words, the phase velocity of the waveguide mode is the same as a plane TEM wave in the same medium.

Now, we assume that a TEM wave exists in both inhomogeneous regions of the microstrip line or all three dielectric regions of the optical fiber in Figure 24.6. Then the phase velocities in the z direction, determined by ω/β_z of each region will be ω/β_i of the respective region. As a consequence, phase matching is not possible, and the boundary condition cannot be satisfied at the dielectric interfaces.

Nevertheless, the lumped element circuit model of the transmission line is still a very good model for such a waveguide. If the line capacitance and line inductance of such lines can be estimated, β_z can still be estimated. As has been shown before, circuit theory is valid when the frequency is low, or the wavelength is large compared to the size of the structures.

24.2.2 Hybrid Modes–Inhomogeneously-Filled Waveguides

It turns out that when such inhomogeneity is present, as in the case of the optical fiber, both TE and TM waves are needed in the waveguide. This will allow the boundary conditions to be satisfied at the interface of different regions [157]. These modes are called hybrid modes. Sometimes, these hybrid modes are called EH or HE modes, as in an optical fiber. Nevertheless, the guidance is still via a bouncing wave picture, where the bouncing waves are reflected off the boundaries of the waveguides. In the case of an optical fiber or a dielectric waveguide, the reflection is due to total internal reflection. But in the case of a PEC metalic waveguides, the reflection is due to the PEC metal walls which is a perfect reflector.

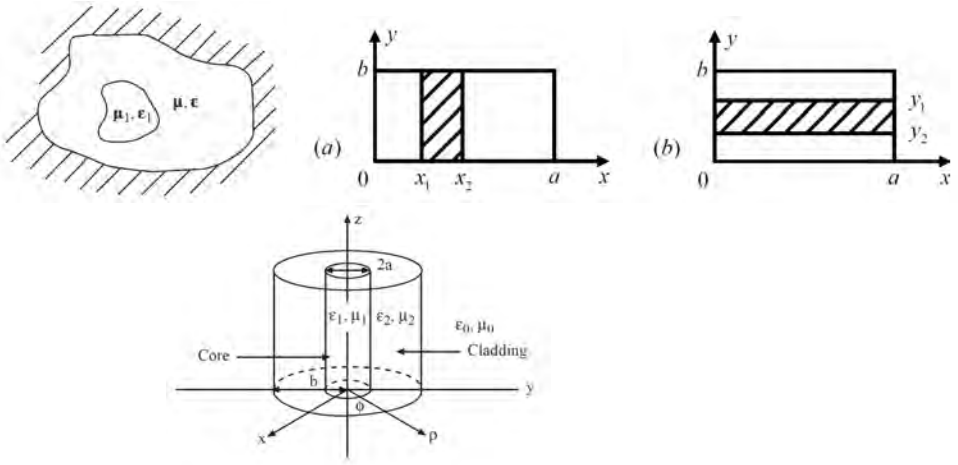


Figure 24.7: Some examples of inhomogeneously filled waveguides where hybrid modes exist: (top-left) A general inhomogeneously filled waveguide, (top-right) two examples of slab-loaded rectangular waveguides, and (bottom) an optical fiber with core and cladding.

24.2.3 Guidance of Modes

Propagation of a plane wave in free space is by the exchange of electric stored energy and magnetic stored energy. So the same physics happens in a waveguide. For example. in the transmission line, the guidance is by the exchange of electric and magnetic stored energy via the coupling between the line capacitance and the line inductance of the line. In this case, the waveguide size, like the cross-section of a coaxial cable, can be made much smaller than the wavelength and the wave is still guided.

In the case of hollow waveguides, the **E** and **H** fields are coupled through their space and time variations representing a bouncing wave inside the waveguide. Namely,

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H}, \qquad \nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E} \tag{24.2.5}$$

Hence, in the frequency domain, the fields are coupled through their spatial derivative, and hence, the exchange of the energies stored is via the space that stores these energies, like that of a plane wave. These waveguides work only when these plane waves can "enter" the waveguide. Therefore, the size of these waveguides has to be about half a wavelength.

The surface plasmonic waveguide is an exception in that the exchange is between the electric field energy stored with the *kinetic energy* stored in the moving electrons in the plasma instead of magnetic energy stored. This form of energy stored is sometimes referred to as coming from *kinetic inductance*. Therefore, the dimension of the waveguide can be very small compared to wavelength, and yet the surface plasmonic mode can be guided.

24.3 "Homomorphism" of Hollow Waveguides and Transmission Lines

Previously, we have demonstrated mathematical "homomorphism" between plane waves in layered medium and transmission lines. Such "homomorphism" can be further extended to hollow waveguides and transmission lines.[6] But unlike the plane wave in layered medium case, we cannot replace the ∇ operator with $-j\boldsymbol{\beta}$ in a hollow waveguide. Hence, the mathematics is slightly more elaborate. We can show this first for TE modes in a hollow waveguide, and the case for TM modes can be established by invoking duality principle.[7]

24.3.1 TE Case

For this case, $E_z = 0$, and from Maxwell's equations

$$\nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E} \tag{24.3.1}$$

First we let $\nabla = \nabla_s + \nabla_z$, $\mathbf{H} = \mathbf{H}_s + \mathbf{H}_z$ where $\nabla_z = \hat{z}\frac{\partial}{\partial z}$, and $\mathbf{H}_z = \hat{z}H_z$, and the subscript s implies transverse to the z components. Then

$$(\nabla_s + \nabla_z) \times (\mathbf{H}_s + \mathbf{H}_z) = \nabla_s \times \mathbf{H}_s + \nabla_z \times \mathbf{H}_s + \nabla_s \times \mathbf{H}_z = j\omega\varepsilon\mathbf{E} \tag{24.3.2}$$

where it is understood that $\nabla_z \times \mathbf{H}_z = 0$. Notice that the first term on the right-hand side of the above is pointing in the z direction. By letting $\mathbf{E} = \mathbf{E}_s + \mathbf{E}_z$, and equating transverse components in (24.3.1), we have[8]

$$\nabla_z \times \mathbf{H}_s + \nabla_s \times \mathbf{H}_z = j\omega\varepsilon\mathbf{E}_s \tag{24.3.3}$$

To simplify further the above equation, we shall relate $\mathbf{H}_z$ from the above with the other field components. To this end, we look at Faraday's law from which we have

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{24.3.4}$$

Again, by letting $\mathbf{E} = \mathbf{E}_s + \mathbf{E}_z$, we can let (24.3.4) be written as

$$\nabla_s \times \mathbf{E}_s + \nabla_z \times \mathbf{E}_s + \nabla_s \times \mathbf{E}_z = -j\omega\mu(\mathbf{H}_s + \mathbf{H}_z) \tag{24.3.5}$$

Equating z components of the above, we have

$$\nabla_s \times \mathbf{E}_s = -j\omega\mu\mathbf{H}_z \tag{24.3.6}$$

The above allows us to express $\mathbf{H}_z$ in terms of $\mathbf{E}_s$. Thus Eq.(24.3.3) can be rewritten as

$$\nabla_z \times \mathbf{H}_s + \nabla_s \times \frac{1}{-j\omega\mu}\nabla_s \times \mathbf{E}_s = +j\omega\varepsilon\mathbf{E}_s \tag{24.3.7}$$

[6] For a waveguide with a center conductor like the coaxial cable, the fundamental mode with no cut-off is the TEM mode, and it actually is the tranmission line mode. The higher-order modes in a coax can be classified into TE and TM modes just like a hollow waveguide.

[7] I have not seen exposition of such mathematical homomorphism elsewhere except in very simple cases [34].

[8] And from the above, it is obvious that $\nabla_s \times \mathbf{H}_s = j\omega\varepsilon\mathbf{E}_z$, but this equation will not be used in the subsequent derivation.

The above can be further simplified by noting that

$$\nabla_s \times \nabla_s \times \mathbf{E}_s = \nabla_s(\nabla_s \cdot \mathbf{E}_s) - \nabla_s \cdot \nabla_s \mathbf{E}_s = -\nabla_s^2 \mathbf{E}_s \tag{24.3.8}$$

But since $\nabla \cdot \mathbf{E} = 0$, and $E_z = 0$ for TE modes, it also implies that $\nabla_s \cdot \mathbf{E}_s = 0$ which justify the last equality in the above. From Maxwell's equations, we have previously shown that for a homogeneous source-free medium,

$$(\nabla^2 + \beta^2)\mathbf{E} = 0, \qquad \text{or} \qquad (\nabla^2 + \beta^2)\mathbf{E}_s = 0 \tag{24.3.9}$$

since $E_z = 0$ for TE mode. Furthermore, assuming $e^{\mp j\beta_z z}$ for the z dependence of the waveguide modes, (24.3.9) then becomes

$$({\nabla_s}^2 + \beta^2 - {\beta_z}^2)\mathbf{E}_s = 0 \tag{24.3.10}$$

or that $\mathbf{E}_s$ satisfies the reduced wave equation. Thus,

$$({\nabla_s}^2 + {\beta_s}^2)\mathbf{E}_s = 0 \tag{24.3.11}$$

where $\beta_s^2 = \beta^2 - \beta_z^2$ is the transverse wave number. Consequently, from (24.3.8), we arrive at the simplification, or that

$$\nabla_s \times \nabla_s \times \mathbf{E}_s = -\nabla_s^2 \mathbf{E}_s = \beta_s^2 \mathbf{E}_s \tag{24.3.12}$$

As such, using this in (24.3.7), it becomes

$$\begin{aligned} \nabla_z \times \mathbf{H}_s &= j\omega\varepsilon \mathbf{E}_s + \frac{1}{j\omega\mu}{\beta_s}^2 \mathbf{E}_s \\ &= j\omega\varepsilon \left(1 - \frac{{\beta_s}^2}{\beta^2}\right) \mathbf{E}_s = j\omega\varepsilon \frac{{\beta_z}^2}{\beta^2} \mathbf{E}_s \end{aligned} \tag{24.3.13}$$

Letting $\beta_z = \beta \cos\theta$, then the above can further be rewritten as

$$\nabla_z \times \mathbf{H}_s = j\omega\varepsilon \cos^2\theta \mathbf{E}_s \tag{24.3.14}$$

Now, the above resembles one of the two telegrapher's equations that we seek.

Again, looking at (24.3.4), assuming $E_z = 0$, equating transverse components, we have

$$\nabla_z \times \mathbf{E}_s = -j\omega\mu \mathbf{H}_s \tag{24.3.15}$$

More explicitly, we can rewrite (24.3.14) and (24.3.15) in the above as

$$\frac{\partial}{\partial z}\hat{z} \times \mathbf{H}_s = j\omega\varepsilon \cos^2\theta \mathbf{E}_s \tag{24.3.16}$$

$$\frac{\partial}{\partial z}\hat{z} \times \mathbf{E}_s = -j\omega\mu \mathbf{H}_s \tag{24.3.17}$$

Visibly, the above resembles the telegrapher's equations. We can multiply (24.3.17) by $\hat{z}\times$ to get

$$\frac{\partial}{\partial z}\mathbf{E}_s = j\omega\mu\hat{z}\times\mathbf{H}_s \tag{24.3.18}$$

Now (24.3.16) and (24.3.18) are a set of coupled equations that look even more like the telegrapher's equations. We can have $\mathbf{E}_s \to V$, $\hat{z}\times\mathbf{H}_s \to -I$, $\mu \to L$, $\varepsilon\cos^2\theta \to C$, and the above resembles the telegrapher's equations, or that the waveguide problem is homomorphic to the transmission line problem. The characteristic impedance of this equivalent line is then

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{\mu}{\varepsilon\cos^2\theta}} = \sqrt{\frac{\mu}{\varepsilon}}\frac{1}{\cos\theta} = \frac{\omega\mu}{\beta_z} \tag{24.3.19}$$

which is the wave impedance of a hollow waveguide. Consequently, the TE modes of a waveguide can be mapped into a transmission problem. This can be done, for instance, for the TE_{mn} mode of a rectangular waveguide. Then, in the above

$$\beta_z = \sqrt{\beta^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \tag{24.3.20}$$

Here, each TE_{mn} mode will be represented by a different equivalent characteristic impedance Z_0, since β_z is different for different TE_{mn} modes.

When this is used in practice, and when multi-modes are present the number of equivalent transmission line can become unwieldy. Engineers like to use simpler systems: Simplicity rules! A simpler alternative is usually sought, such as the mode-matching method [157].

24.3.2 TM Case

This case can be derived using duality principle. Invoking duality, and after some algebra, then the equivalence of (24.3.16) and (24.3.18) become

$$\frac{\partial}{\partial z}\mathbf{E}_s = j\omega\mu\cos^2\theta\hat{z}\times\mathbf{H}_s \tag{24.3.21}$$

$$\frac{\partial}{\partial z}\hat{z}\times\mathbf{H}_s = j\omega\varepsilon\mathbf{E}_s \tag{24.3.22}$$

To keep the dimensions commensurate, we let $\mathbf{E}_s \to V$, $\hat{z}\times\mathbf{H}_s \to -I$, $\mu\cos^2\theta \to L$, $\varepsilon \to C$. Again the above resembles the telegrapher's equations. We can thus let

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{\mu\cos^2\theta}{\varepsilon}} = \sqrt{\frac{\mu}{\varepsilon}}\cos\theta = \frac{\beta_z}{\omega\varepsilon} \tag{24.3.23}$$

Please note that (24.3.19) and (24.3.23) are very similar to that for the plane wave case, which are the wave impedance for the TE and TM modes, respectively.

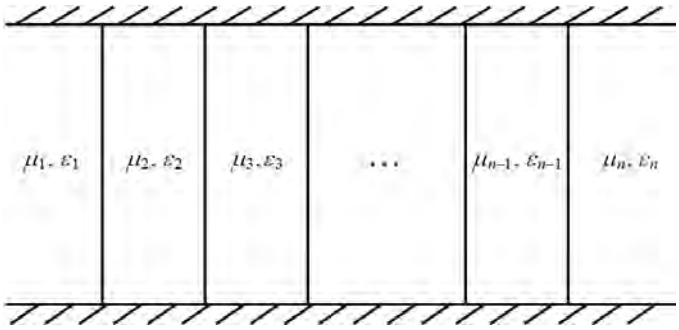


Figure 24.8: A waveguide filled with layered medium is mathematically homomorphic to a multi-section transmission line problem. In this problem, there is no mode conversion, and hence, a single mode can propagate through the waveguide and yet, the boundary conditions at the interfaces can be satisfied, like plane waves in layered media. Hence, transmission-line method can be used to solve this problem.

The above implies that if we have a waveguide of arbitrary cross section filled with layered media, the problem can be mapped to a multi-section transmission line problem. Then it can be solved with transmission line methods. When V and I are continuous at a transmission line junction, $\mathbf{E}_s$ and $\mathbf{H}_s$ will also be continuous. Therefore, the transmission line solution would also imply continuous $\mathbf{E}_s$ and $\mathbf{H}_s$ field solutions. No mode conversion happens at such junctions, as shall be explained later.

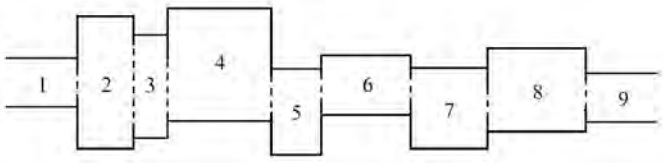


Figure 24.9: A multi-section waveguide is not exactly homormorphic to a multi-section transmission line problem, when the cross section of the waveguides are not equal to each other. The boundary conditions at the junction cannot be met by assuming a single mode. Therefore, multi-modes are assumed and needed to match the boundary conditions [157]. Circuit elements are needed at the junctions to approximately capture the physics at the waveguide junctions as shown in the next figure.

24.3.3 Mode Conversion

In the waveguide shown in Figure 24.8, there is no mode conversion at the junction interface. Assuming a rectangular waveguide as an example, what this means is that if we send a TE_{10} into the waveguide, this same mode will propagate throughout the length of the waveguide. The reason

is that only this mode alone is sufficient to satisfy the boundary condition at the junction interface. The mode profile does not change throughout the length of the waveguide.

To elaborate further, from our prior knowledge, the transverse fields of the waveguide, e.g., for the TM mode, can be derived to be

$$\mathbf{H}_s = \nabla \times \hat{z}\Psi_{es}(\mathbf{r}_s)e^{\mp j\beta_z z} \tag{24.3.24}$$

$$\mathbf{E}_s = \frac{\mp\beta_z}{\omega\varepsilon}\nabla_s\Psi_{es}(\mathbf{r}_s)e^{\mp j\beta_z z} \tag{24.3.25}$$

In the above, β_s^2 and $\Psi_{es}(\mathbf{r}_s)$ are eigenvalue and eigenfunction, respectively, that depend only on the geometrical cross-sectional shape of the waveguide, but not the materials filling the waveguide. These eigenfunctions are the same throughout different sections of the waveguide and only $\beta_z = \sqrt{\beta_i^2 - \beta_s^2}$ changes from section to section. Therefore, with the incident mode at the junction, the boundary conditions can be easily satisfied at the junctions with the inclusion of reflected and transmitted waves of the same mode.

However, for a multi-junction waveguide show in Figure 24.9, tangential **E** and **H** continuous condition cannot be satisfied by a single mode in each waveguide alone: V and I continuous at a transmission line junction will not guarantee the continuity of tangential **E** and tangential **H** fields at the waveguide junction.

Multi-modes have to be assumed on both sides of the junction at each section in order to match boundary conditions at the junction [90]. Moreover, mode matching method for multiple modes has to be used at each junction. Typically, a single mode incident at a junction will give rise to multiple modes reflected and multiple modes transmitted. The multiple modes give rise to the phenomenon of *mode conversion* at a junction. Hence, the waveguide may need to be modeled with multiple transmission lines where each mode is modeled by a different transmission line with different characteristic impedances. Having to model a waveguide junction with multiple transmission line is unwieldy, and is usually avoided. Instead, mode matching method is used to solve this problem [157].

However, the operating frequency can be chosen so that only one mode is propagating at each section of the waveguide, and the other modes are cut-off or evanescent. Then one can assume a single-mode approximation of the waveguide modes in each section of the waveguide. In this case, the higher-order modes are evanescent away from the waveguide junction, giving rise to localized energy storage at a junction in the **E** and **H** fields. These energies can be either inductive or capacitive. The junction effects may be modeled by a simple circuit model as shown in Figure 24.10. These junction elements also account for the physics that the currents and voltages are not continuous anymore across the junction. Moreover, these junction lumped circuit elements account for the stored electric and magnetic energies at the junction.

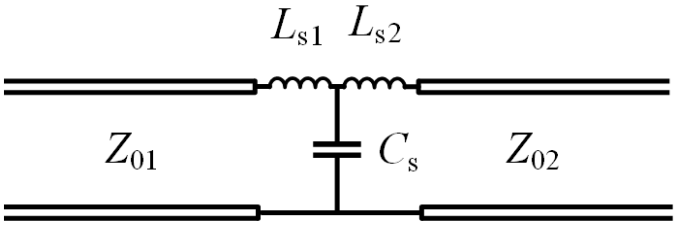


Figure 24.10: When the single-mode approximation holds, one assumes that away from the waveguide junction, there is only one single mode. Close to the waveguide junction, however, the evanescent higher order modes, give rise to localize stored electric and magnetic energies at the waveguide junction. Then, circuit elements are used to account for these stored energies at the waveguide junctions. They also account for that the currents and voltages are not continuous across the junctions anymore as the fields of the dominant modes in each section as shown in Figure 24.9 are not continuous anymore. The moral of the story is that engineers love to replace complicated theory with simple ones in order to solve complex problems. Simplicity rules!

Exercises for Lecture 24

Problem 24-1:

(i) Explain when you would have quasi-TEM modes and hybrid modes.

(ii) Show that the rectangular or circular waveguide problem can be made homomorphic to a transmission line problem.

(iii) Explain how you would map the waveguide parameters to transmission line parameters for the TE and TM modes of a hollow waveguide.

(iv) Derive (24.3.24) and (24.3.24). Explain why mode conversion occurs at a waveguide junction.

Chapter 25

Cavity Resonators

Cavity resonators are important components of microwave and optical systems. They work by constructive and destructive interference of bouncing waves in an enclosed region. Just as LC resonators in circuits, they can be used as filters, or as devices to enhance certain physical interactions. These can happen in radiation antennas or electromagnetic sources such as magnetrons or lasers. They can also be used to make high sensitivity sensors. We will study a number of them. For many of them, we will discuss them only heuristically in this lecture. Their full solutions are usually obtained using numerical methods.

25.1 Transmission Line Model of a Resonator

The simplest cavity resonator is formed by using a transmission line. The source end can be terminated by Z_S and the load end can be terminated by Z_L. When Z_S and Z_L are non-dissipative, such as when they are reactive loads (capacitive or inductive), the reflection coefficient from (16.1.9), for $Z_L = jX$ where X is the reactance of the load, is

$$\Gamma_L = \frac{jX - Z_0}{jX + Z_0} \tag{25.1.1}$$

Like the Fresnel reflection coefficient for the total internal reflection case as shown in (18.2.3), it has a magnitude of one indicating that the energy is totally reflected. Thus no energy is dissipitated as a wave is totally reflected off them. Therefore, if the wave can bounce and interfere constructively between the two ends, a coherent solution or a resonant solution can exist due to constructive inference.

The resonant solution exists even when the source is turned off. In mathematical parlance, this is a homogeneous solution to a partial differential equation or ordinary differential equation, since the right-hand side of the pertinent equation is zero. The right-hand side of these equations usually corresponds to a source term or a driving term. In physics parlance, this is a natural solution since it exists naturally without the need for a driving or exciting source. In linear algebra, when the matrix equation $\overline{\mathbf{A}} \cdot \mathbf{x} = 0$ has a non-trivial solution, it is the null-space solution.

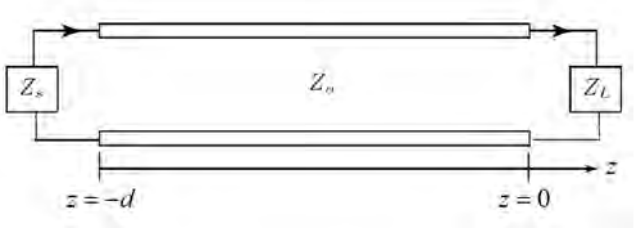


Figure 25.1: A simple ideal resonator is made by terminating a transmission line with two reactive loads at its two ends, the source end with Z_S and the load end with Z_L.

The transverse resonance condition for 1D problem can be used to derive the resonance frequency, namely that

$$1 = \Gamma_S \Gamma_L e^{-2j\beta_z d} \tag{25.1.2}$$

where Γ_S and Γ_L are the reflection coefficients at the source and the load ends, respectively, β_z the the wave number of the wave traveling in the z direction, and d is the length of the transmission line.

For a TEM mode in the transmission line, as in a coax filled with homogeneous medium, then $\beta_z = \beta$, where β is the wavenumber for the homogeneous medium. Otherwise, for a quasi-TEM mode, $\beta_z = \beta_e$ where β_e is some effective wavenumber for a z-propagating wave in an inhomogeneous (heterogeneous) medium. In general,

$$\beta_e = \omega / v_e \tag{25.1.3}$$

where v_e is the effective phase velocity of the wave in the heterogeneous structure like a microstrip line.

When the source and load impedances are replaced by short or open circuits, then the reflection coefficients are simply -1 for a short, and $+1$ for an open circuit. The (25.1.2) above then becomes

$$\pm 1 = e^{-2j\beta_e d} \tag{25.1.4}$$

The $\pm$ signs correspond to different combinations of open and short circuits at the two ends of the transmission lines. When a "+" sign is chosen, it corresponds to either both ends are short circuit, or are open circuit. Then the resonance condition is such that

$$\beta_e d = p\pi, \quad p = 0, 1, 2, \ldots, \quad \text{or integer} \tag{25.1.5}$$

For a TEM or a quasi-TEM mode in a transmission line, $p = 0$ is not allowed as the voltage is constant, and it will be uniformly zero on the transmission line. (If only $V(z) = 0$ at one end, it will be zero for all z implying a trivial solution.) The lowest mode then is when $p = 1$ corresponding to a half wavelength on the transmission line. The voltage distribution is shown in Figure 25.2.

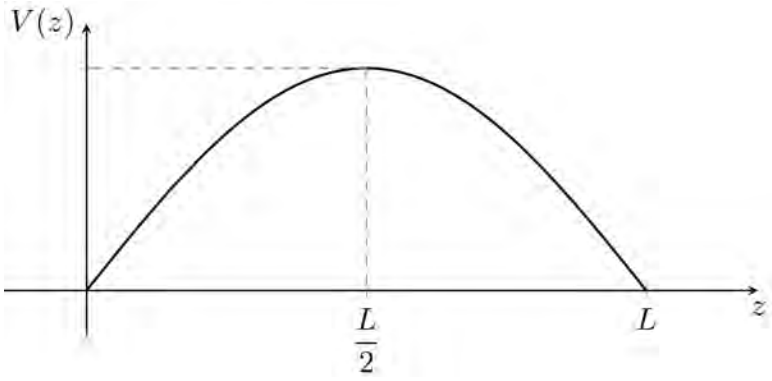


Figure 25.2: The voltage distribution on a half-wave transmission line resonator at its resonant frequency. Since it needs a half wavelength to resonate, it is not the smallest resonator, and hence, it is not employed in cell-phone antennas.

When the line is open at one end, and shorted at the other end in (25.1.2), the resonance condition corresponds to the "−" sign in (25.1.4), which gives rise to

$$e^{-2j\beta_e d} = e^{-jp\pi} = -1, \quad p \quad \text{odd integer} \tag{25.1.6}$$

The above implies that

$$\beta_e d = p\pi/2, \quad p \quad \text{odd integer} \tag{25.1.7}$$

As shown in Figure 25.3, the lowest mode is when $p = 1$ corresponding to a quarter wavelength on the transmission line, which is smaller than that of a half-wavelength transmission line terminated with short or open at both ends. Designing a small resonator using a quarter-wave resonator is a prerogative in modern day electronic design. For example, miniaturization in cell phones calls for smaller components that can be packed into smaller spaces.

A quarter wavelength resonator has a voltage distribution shown in Figure 25.3. One that is made with a with a coax is shown in Figure 25.4. It is easier to make a short approximately as indicated at the left end: a good metallic conductor like copper suffices because it close to a PEC at microwave frequency.[1] But it is hard to make a true open circuit as shown at the right end. A true open circuit means that the current has to be zero. But when a coax is terminated with an open, the electric current does not end abruptly. The fringing field at the right end gives rise to stray capacitance through which displacement current can flow in accordance to the generalized Ampere's law. Hence, we have to model the right end termination with a small stray or fringing field capacitance as shown in Figure 25.4. This makes the transmission line slightly longer.

To design a true open circuit, one needs to terminate the right end of the transmission line with a perfect magnetic conductor (PMC) in theory. By going through *Gedanken* experiment,

[1] A PEC has a skin depth of zero. A good conductor like copper will have skin depth much less than the wavelength at microwave frequencies, and hence, it behaves essentially like a PEC.

since $H_\phi = 0$ at the termination, one can show that the current at the right termination has to be zero.

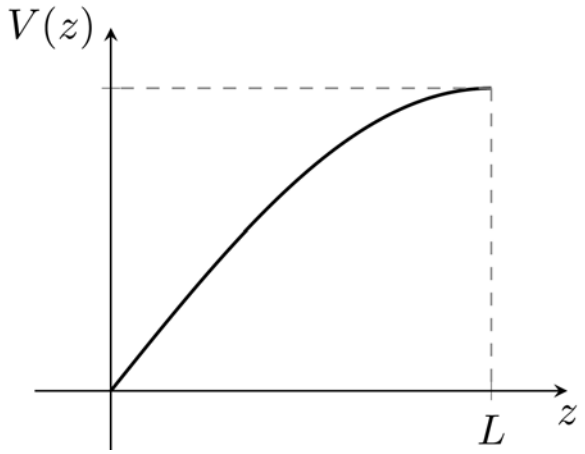


Figure 25.3: The voltage distribution on a quarter-wave transmission line resonator at its lowest resonant mode. Since it needs only a quarter wavelength to resonate, it can be made small compared to wavelength, and hence, it is a very popular resonator used in cell-phone antennas.

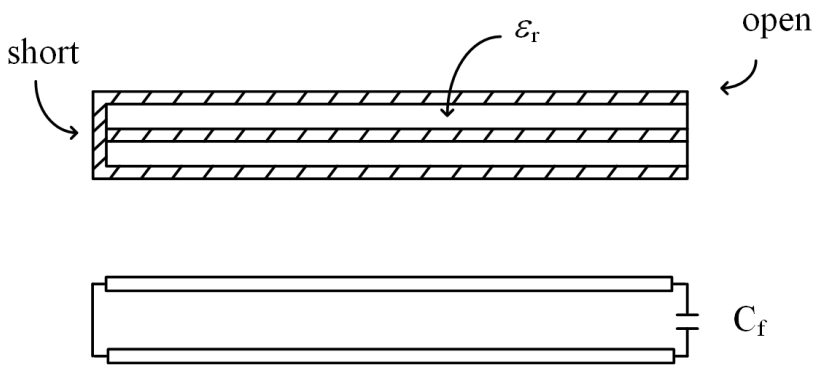


Figure 25.4: A short and open circuited transmission line can be a resonator, but the open end has to be modeled with a fringing field capacitance C_f since there is no exact or true open circuit. The resonance condition will have to be derived from (25.1.2), which will usually give a transcendental equation. Graphical method can be used to solve the transcendental equation.

25.2 Cylindrical Waveguide Resonators

Since a cylindrical waveguide[2] is "homomorphic" to a transmission line, we can model a mode in this waveguide as a transmission line. Then the termination of the waveguide with either a short or an open circuit at its end makes it into a resonator.

Again, there is no true open circuit in an open ended waveguide, as there will be fringing fields at its open ends. If the aperture is large enough, the open end of the waveguide radiates and may be used as an antenna as shown in Figure 25.5.

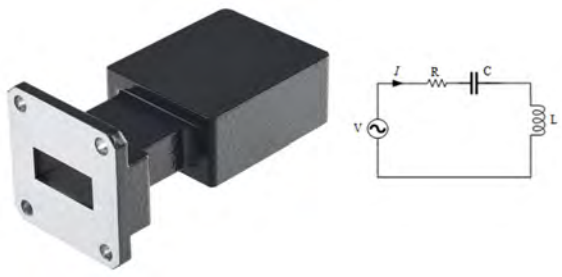


Figure 25.5: A rectangular waveguide terminated with a short at one end, and an open circuit at the other end. The open end can also act as an antenna as it radiates. When the cavity is injected with electromagnetic fields coinciding with its resonance frequency, the fields inside the cavity becomes large, so does the fields at the aperture, making it a better radiator. This is a cavity-backed antenna: it uses resonance tunneling to enhance it radiation capability. (Resonance tunneling phenomenon describes the ease of photon to tunnel through a barrier when the photon is close to the resonance frequency of the barrier structure.). The RLC circuit can approximate the physics of the structure (courtesy of RFcurrent.com).

25.2.1 Rectangular Cavity Resonator

As previously shown, single-section waveguide resonators can be modeled with a transmission line using "homomorphism" with the appropriately chosen β_z. Then, $\beta_z = \sqrt{\beta^2 - \beta_s^2}$ where β_s can be found by first solving a 2D waveguide problem corresponding to the reduced-wave equation.

For a rectangular waveguide, for example, from the previous lecture,

$$\beta_z = \sqrt{\beta^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \tag{25.2.1}$$

for both TE_{mn} and TM_{mn} modes.[3] If the waveguide is terminated with two shorts (which is easy to make) at its ends, then the resonance condition is that

$$\beta_z = p\pi/d, \quad p \text{ integer} \tag{25.2.2}$$

[2]Both rectangular and circular waveguides are cylindrical waveguides.

[3]It is noted that for a certain mn mode, with a choice of frequency, $\beta_z = 0$ which does not happen in a transmission line.

Together, using (25.2.1), we have the condition that

$$\beta^2 = \frac{\omega^2}{c^2} = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{p\pi}{d}\right)^2 \tag{25.2.3}$$

The above can only be satisfied by certain select frequencies: these frequencies are the resonant frequencies of the rectangular cavity. The corresponding mode is called the TE_{mnp} mode or the TM_{mnp} mode depending on if these modes are TE to z or TM to z. One can think of these modes as a consequence of the TE_{mn} or TM_{mn} modes in the rectangular waveguide bouncing back and forth in the z direction.

The entire electromagnetic fields of the cavity can be found from the pilot scalar potentials previously defined, namely that

$$\mathbf{E} = \nabla \times \hat{z}\Psi_h, \quad \mathbf{H} = \nabla \times \mathbf{E}/(-j\omega) \tag{25.2.4}$$

$$\mathbf{H} = \nabla \times \hat{z}\Psi_e, \quad \mathbf{E} = \nabla \times \mathbf{H}/(j\omega\varepsilon) \tag{25.2.5}$$

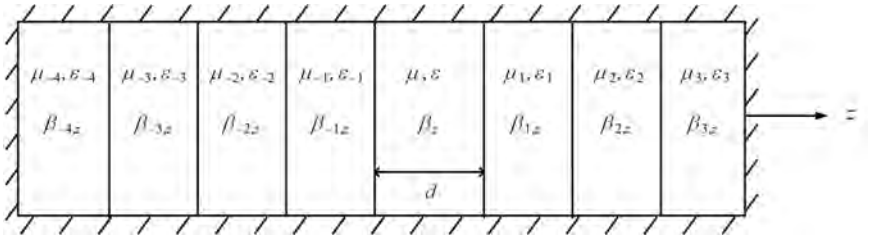


Figure 25.6: A waveguide filled with layered dielectrics can also become a resonator. The transverse resonance condition can be used to find the resonant modes. This can be obtained by exploiting the mathematical "homomorphism" between the waveguide problem and the transmission line problem.

25.2.2 Layered Medium Cavity

Since the layered medium problem in a waveguide is the same as the layered medium problem in open space, we can use the generalized transverse resonance condition to find the resonant frequencies and hence the modes of a waveguide cavity loaded with layered medium as shown in Figure 25.6. This condition is repeated below as:

$$\tilde{R}_- \tilde{R}_+ e^{-2j\beta_z d} = 1 \tag{25.2.6}$$

where d is the length of the waveguide section and the above condition is applied. Here, $\tilde{R}_-$ and $\tilde{R}_+$ are the generalized reflection coefficients to the left and right of the center waveguide section. The above is similar to the resonant condition using the transmission line model in equation (25.1.2), except that now, we have replaced the transmission line reflection coefficient with TE or TM generalized reflection coefficients.

25.2.3 Lowest Mode of a Rectangular Cavity

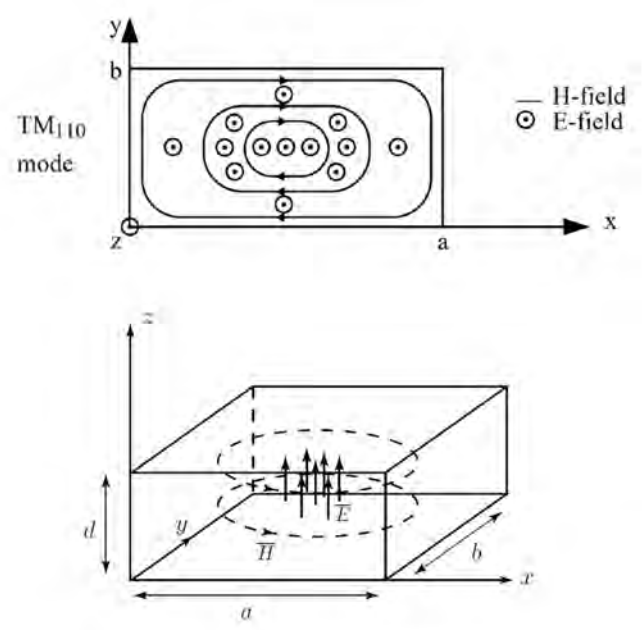


Figure 25.7: The top and 3D views of the E and H fields of the TM_{110} mode of a rectangular resonant cavity. Since this is a snapshot view of two sinusoidal fields that are 90 degrees out of phase, the relative signs of E and H fields are immaterial.

To find the lowest mode of a rectangular cavity is quite tricky. We assume that $a > b > c$, and find a mode where the combination of m, n, p in (25.2.3) will make β the smallest. We can first try to make one of them zero so that β will be greatly reduced. It is not possible to make both of them zero as this will reduce it to a 1D problem. We begin by assuming a guided TM to z mode bouncing between two ends of a rectangular cavity. The lowest TM mode is the TM_{11} mode. At the cut-off of this mode, the $\beta_z = 0$ or $p = 0$, implying no variation of the field in the z direction. Let us call this mode the TM_{110} mode. The fields of this mode are visualized in Figure 25.7. It is a mode that bounces in the x and y directions, but does not propagate in the z direction. The electric field is pointing in the z direction in the cavity. Now, we do a Gedanken experiment of putting two metallic shorts above and below the cavity. The boundary conditions at these two end caps are still satisfied since tangential $\mathbf{E}$ field is zero. Hence, this mode can exist and satisfy the boundary conditions on the six walls of the cavity with one of β_i, $i = x, y, z$ being zero. This is the lowest resonant mode, that makes the right-hand side of (25.2.3) as small as possible by setting $p = 0$.

The top and side views of the fields of this mode is shown in Figure 25.7. The corresponding

resonant frequency of this mode, from (25.2.3), satisfies the equation

$$\frac{\omega_{110}^2}{c^2} = \left(\frac{\pi}{a}\right)^2 + \left(\frac{\pi}{b}\right)^2 \tag{25.2.7}$$

Looking at the TE to z modes more carefully, it is required that $p \neq 0$, otherwise, the field is zero in the cavity. For example, it is possible to have the TE_{101} mode with nonzero $\mathbf{E}$ field. The resonant frequency of this mode, from (25.2.3), is

$$\frac{\omega_{101}^2}{c^2} = \left(\frac{\pi}{a}\right)^2 + \left(\frac{\pi}{d}\right)^2 \tag{25.2.8}$$

Clearly, this mode has a higher resonant frequency compared to the TM_{110} mode if $d < b$.

25.2.4 Circular, Cylindrical, and Spherical Cavity Cases

The above analysis can be applied to circular and other cylindrical waveguides with β_s determined differently. For instance, for a circular waveguide, β_s is determined differently using Bessel functions, and for a general arbitrarily shaped waveguide as shown in Figure 25.9, β_s may have to be determined numerically.

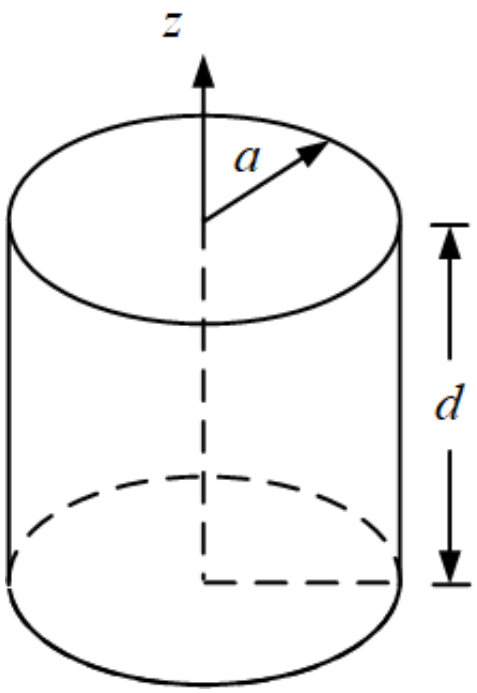


Figure 25.8: A circular resonant cavity made by terminating a circular waveguide at both ends (courtesy of Kong [34]).

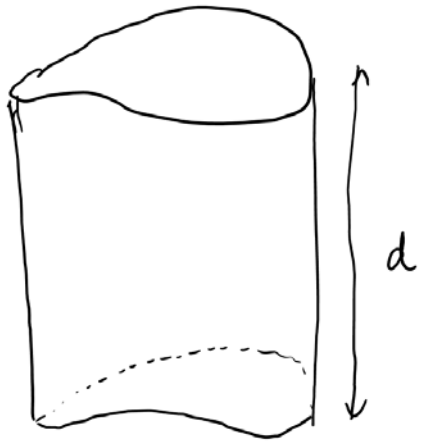


Figure 25.9: An arbitrary cylinder resonant cavity made by terminating an arbitrary waveguide at both ends (courtesy of Kong.

For a spherical cavity, one would have to analyze the problem in spherical coordinates. The equations will have to be solved by the separation of variables using spherical harmonics. Details are given on p. 468 of Kong [34]. These days, when the cavity is of arbitrary shape, numerical methods can be used to find its resonant frequencies.

25.2.5 A General 3D Cavity

For a source-free cavity, the electromagnetics fields inside the cavity satisfies

$$(\nabla^2 + \omega\mu\varepsilon)\mathbf{E}(\mathbf{r}) = 0 \tag{25.2.9}$$

25.3 Some Applications of Resonators

Resonators in microwaves and optics can be used for designing filters, energy trapping devices, and antennas. As filters, they are used like LC resonators in circuit theory. A concatenation of them can be used to narrow or broaden the bandwidth of a filter. As an energy trapping device, a resonator can build up a strong field inside the cavity if it is excited with energy close to its resonance frequency similar to an LC tank circuit. They can be used in klystrons and magnetrons as microwave sources, also as a laser cavity for optical sources, or as a wavemeter to measure the frequency of the electromagnetic field at microwave frequencies. An antenna is a radiator that we will discuss more fully later. The use of a resonator can help in resonance tunneling to enhance the radiation efficiency of an antenna.

25.3.1 Filters

Circuit theory plays an important role in the design of microwave filters, as circuit theory is simple. An LC tank circuit can be used as a simple filter in electronic circuits. A concatenation of a number of LC tank circuits can be used to design a broadband filter. By the same token, microstrip line resonators, and a concatenation of them, are often used to make filters [171].

Transmission lines are often used to model microstrip lines in a complex microwave integrated circuits (MIC) or monolithic MIC (MMIC). In these circuits, due to the etching process, it is a lot easier to make an open circuit rather than a short circuit. But a true open circuit is hard to make as an open ended microstrip line has fringing field at its end as shown in Figure 25.10 [172, 173]. The fringing field gives rise to fringing field capacitance as shown in Figure 25.4. Then the appropriate Γ_S and Γ_L can be used to model the effect of fringing field capacitance. Figure 25.11 shows a concatenation of several microstrip resonators to make a microstrip filter. This is like using a concatenation of LC tank circuits to design filters in circuit theory.

Optical filters can be made with optical etalon as in a Fabry-Perot resonator, or concatenation of them. This is shown in Figure 25.12.

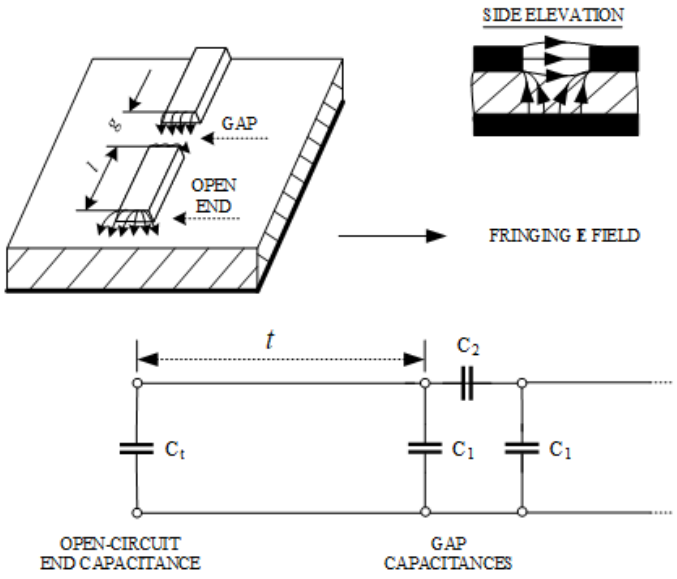


Figure 25.10: End effects and junction effects in a microwave integrated circuit are important [172, 173] (courtesy of Microwave Journal). The fringing fields at the end of a microstrip line or between two microstrip lines can be modeled with parasitic capacitances. The reduction of a complex microwave circuit to a simple lumped-element approximation is a process known as parameter extraction.

Figure 25.11: A microstrip filter designed using concatenated resonators. The connectors to the coax cable are the SMA (sub-miniature type A) connectors (courtesy of aginas.fe.up.pt).

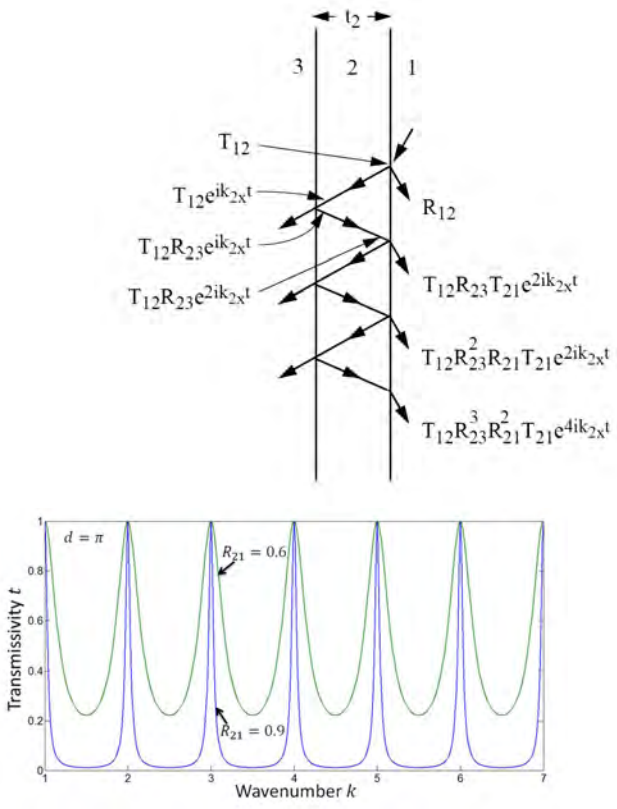


Figure 25.12: Design of a Fabry-Perot resonator [174, 59, 175, 90]. As the magnitude of the reflection coefficient becomes close to one, the wave is better trapped inside the slab. A resonant mode exists inside the slab, providing a means for resonance tunneling.

25.3.2 Electromagnetic Sources (Heuristically)

Microwave sources are often made by transferring kinetic energy from an electron beam to microwave energy. Klystrons, magnetrons, and traveling wave tubes are examples of such devices. However, the cavity resonator in a klystron enhances the interaction of the electrons with the microwave field allowing for more effective energy transfer, causing the field to grow in amplitude as shown in Figure 25.13.

Magnetron cavity works also by transferring the kinetic energy of the electron into the microwave energy. By injecting hot electrons into the magnetron cavity, the electromagnetic cavity resonance is magnified by the absorption of kinetic energy from the hot electrons, giving rise to amplified microwave energy as shown in Figure 25.14.

Figure 25.15 shows laser cavity resonator used to enhance of light wave interaction with material media. By using stimulated emission of electronic transition (the physics is beyond the purview of this course), light energy can be amplified.

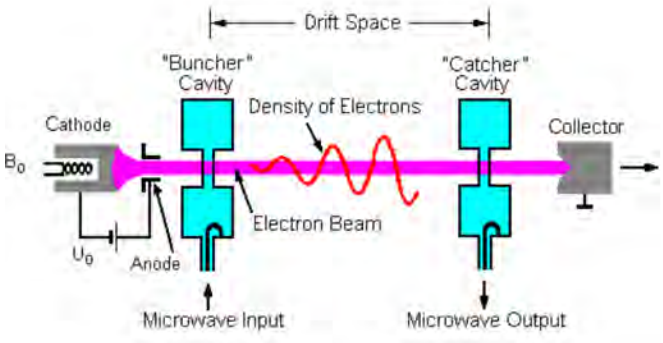


Figure 25.13: A klystron works by converting the kinetic energy of an electron beam into the energy of a traveling microwave next to the beam. As the microwave rides on the electron beam, it absorbs energy from the kinetic energy of the electrons making its amplitude grow as it propagates. The amplified microwave can be collected by the "catcher" cavity (courtesy of Wiki [176]).

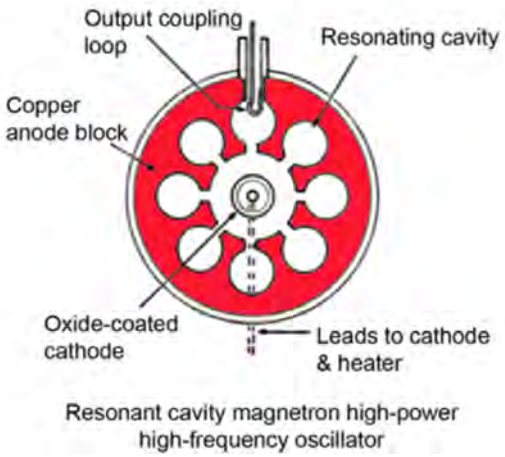


Figure 25.14: A magnetron works by having a high-Q microwave cavity resonator. When the cavity is injected with energetic electrons from the cathode to the anode, the kinetic energy of the electron feeds into the energy of the microwave. The cavity resonance amplifies this field-electron interaction causes energy transfer from the kinetic energy of the electrons to the electromagnetic field energy (courtesy of Wiki [177]).

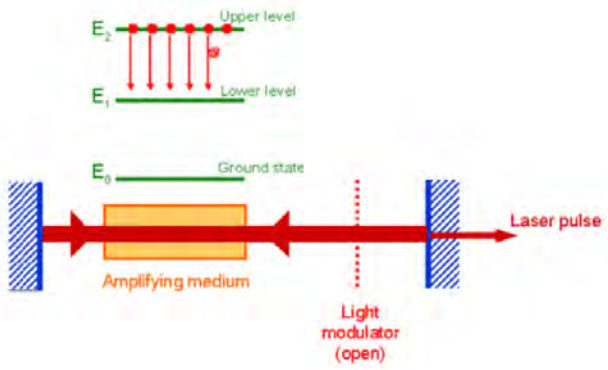


Figure 25.15: A simple view of the physical principle behind the working of the laser. The cavity again enhances the interaction of the photons with the amplifying medium (courtesy of www.optique-ingenieur.org).

Energy trapping of a waveguide or a resonator can be used to enhance the efficiency of a semiconductor laser as shown in Figure 25.16. The trapping of the light energy by the heterojunctions as well as the index profile allows the light to interact more strongly with the lasing medium or

the active medium of the laser. This enables a semiconductor laser to work at room temperature. In 2000, Z. I. Alferov and H. Kroemer, together with J.S. Kilby, were awarded the Nobel Prize for information and communication technology: Alferov and Kroemer for the invention of room-temperature semiconductor laser, and Kilby for the invention of electronic integrated circuit (IC) or the chip.

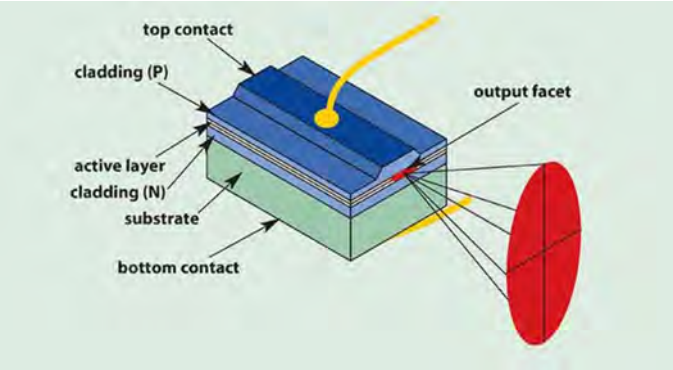


Figure 25.16: A semiconductor laser (also known as laser diode) at work. Room temperature lasing is possible due to both the tight confinement of both light photons and electron-hole pair carriers. Their close proximity causes the diode to lase at room temperature (courtesy of Photonics.com).

25.3.3 Frequency Sensor

A cavity resonator can be used as a frequency sensor. It acts as an energy trap, because it will siphon off energy from a microwave waveguide when the microwave frequency hits the resonance frequency of the cavity resonator. This can be used to determine the frequency of the passing wave. Wavemeters are shown in Figures 25.17 and 25.18. As seen in the picture, there is an entry microwave port for injecting microwave into the cavity, and another exit port for the taking the microwave out of the cavity sensor. The resonant frequency of the cavity can be continuously tuned by changing the location of the plunger. The passing microwave, when it hits the resonance frequency of the cavity, will create a large field inside it due to resonance coupling. The larger field will dissipate more energy on the cavity metallic wall, and gives rise to less energy leaving the cavity. This dip in energy transmission at the resonant freqeuncy of the cavity reveals the frequency of the microwave.

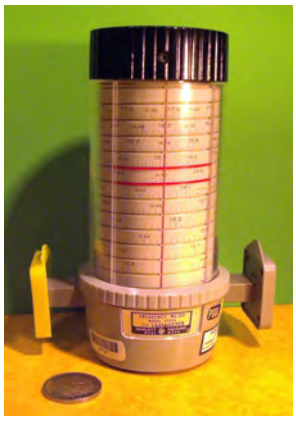


Figure 25.17: An absorption wave meter can be used to measure the frequency of microwave. If the microwave energy enters the cavity at its resonant frequency, strong field buildup inside the cavity causes increased loss and absorption of the microwave energy by the cavity. A dip in energy level of the transmitted signal indicates the coincidence of the resonant frequency of the microwave with the frequency of the passing microwave (courtesy of Wiki [178]).

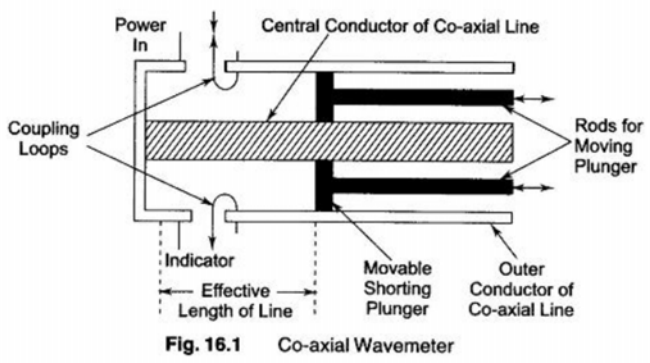


Figure 25.18: The innards of a wavemeter. The location of the plunger short can be continuously moved by rotating the cap of the cavity shown in the previous figure (courtesy of eeeguide.com).

Due to the complexity of waveguides and resonators, most modern waveguide and resonator problems are solved with numerical methods. A good recent review paper is given in [179].

Exercises for Lecture 25

Problem 25-1:

(i) Explain what happens at the junction of two hollow waveguides if they do not share the same cross section.

(ii) Explain why the lowest resonant mode of a rectangular cavity is the TM_{110} mode when $a > b > d$.

(iii) Name some applications of a cavity resonator.

(iv) Using an LC tank circuit model for cavity-backed antenna, explain why the fields are strong inside the cavity when one operates near the resonant frequency of the cavity.

Chapter 26

Quality Factor of Cavities, Mode Orthogonality

Cavity resonators are important for making narrow band filters. The bandwidth of a filter is related to the Q or the quality factor of the cavity. As shown previously, a concatenation of cavity resonators can be used to engineer different filter designs. Resonators can also be used to design various sensing systems, as well as measurement systems. We will study the concept of Q in this lecture.

Also, before we leave the lectures on waveguides and resonators, it will be prudent to discuss mode orthogonality. Since this concept is very similar to eigenvector orthogonality found in matrix or linear algebra, we will relate mode orthogonality in waveguides and cavities to eigenvector orthogonality.

26.1 The Quality Factor of a Cavity—General Concept

The quality factor of a cavity or its Q measures how ideal or lossless a cavity resonator is. An ideal lossless cavity resonator will sustain free oscillations forever, while most resonators sustain free oscillations for a finite time due to loss or damping of the oscillation. This is because of losses coming from radiation, dissipation in the dielectric material filling the cavity, or resistive loss of the metallic part of the cavity. When loss is incurred from radiation, the term "radiation damping" is often used.

26.1.1 Analogue with an LC Tank Circuit

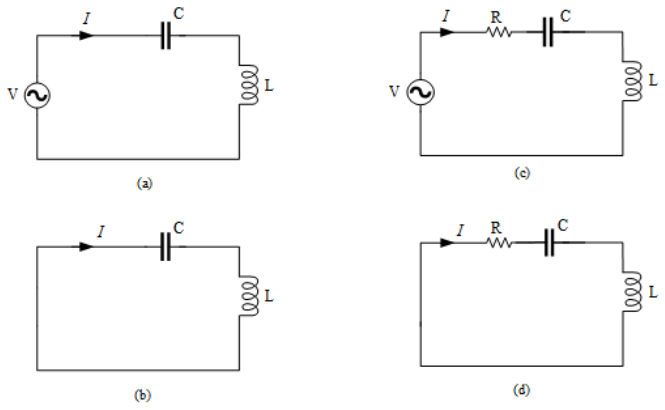


Figure 26.1: For the circuit on the left, it will resonate forever even if the voltage source is turned off. But for the circuit on the right, if the source is turned off, the current in the circuit will decay with time due to dissipation in the resistor and damping due to loss.

Much physical insight can be gotten by studying a simple resonator. One of the simplest resonators imaginable is the LC tank circuit. By using it as an analogue, we can better understand the resonance of a cavity. In theory, when there is no loss in an LC tank circuit, it can oscillate forever. Moreover, if we turn off the source as in Figure 26.1(b), a free oscillation solution persists in the circuit.[1]

One can write the voltage-current relation in the lossless circuit as

$$I(\omega) = \frac{V(\omega)}{j\omega L + 1/(j\omega C)} = V(\omega)Y(\omega) \tag{26.1.1}$$

where

$$Y(\omega) = \frac{1}{j\omega L + 1/(j\omega C)} \tag{26.1.2}$$

The above $Y(\omega)$ in (26.1.2) can be thought of as the transfer function of the linear time-invariant system where the input is $V(\omega)$ and the output is $I(\omega)$. When the voltage is zero or turned off, a non-zero current exists or persists at the resonance frequency of the oscillator. The resonant frequency is when the denominator in the above equation is zero, so that I is finite despite $V = 0$.[2] This resonant frequency, obtained by setting the denominator of Y to zero, is given by $\omega_R = 1/\sqrt{LC}$.

[1] This is analogous to the homogeneous solution of an ordinary differential equation or the natural solution of a physical system.

[2] We take advantage of the fact that zero divided by zero is undefined.

When a small resistor is added in the circuit to give rise to loss, the voltage-current relation becomes

$$I(\omega) = \frac{V(\omega)}{j\omega L + R + 1/(j\omega C)} = V(\omega)Y(\omega) \tag{26.1.3}$$

where

$$Y(\omega) = \frac{1}{j\omega L + R + 1/(j\omega C)} \tag{26.1.4}$$

Now, the denominator of the above functions can never go to zero for real ω. But there exists complex ω that will make Y become infinite. These are the complex resonant frequencies of the circuit. Thus, the homogeneous solution (also called the natural solution, or free oscillation) can only exist at the complex resonant frequencies for this lossy circuit. With complex resonances, the voltage and the current are decaying sinusoids.

To further convince yourself, $Y(\omega)$ is the impulse response of the system which is the response of the system when it is excited with an impulse function. One can easily find the response by taking the inverse Fourier transform of (26.1.4). The inverse Fourier transform of such function, for instance, is found in many textbooks on signals and systems [55] [Table 7.2]. It can be shown then that the response of the system or $i(t)$ which is the inverse Fourier transform of $I(\omega)$ above is of the form

$$i(t) \sim I_0 e^{-\alpha t}\cos(\omega_0 t + \phi)u(t) \tag{26.1.5}$$

where $u(t)$ is the unit step function.

By the same token, the impulse response of a electromagnetic cavity can be represented by poles, each of which represents the resonant mode of the cavity. Each of these modes corresponds to the free oscillation of the cavity. (To simplify matter, we choose the operating frequency in the vicinity of the resonant frequency, and then the impulse response can be approximated by a single pole system.) Because of losses, the free oscillation has a complex frequency given by $\omega_R = \omega_0 + j\alpha$. Upon Fourier inverse transforming, the cavity has electromagnetic field with time dependence as follows:[3]

$$\mathbf{E} \propto e^{-\alpha t}\cos(\omega t + \phi_1), \quad \mathbf{H} \propto e^{-\alpha t}\cos(\omega t + \phi_2) \tag{26.1.6}$$

That is, they are decaying sinusoids. The total stored energy of the system is proportional to $\frac{1}{4}\varepsilon|\mathbf{E}|^2 + \frac{1}{4}\mu|\mathbf{H}|^2$. But by assuming the loss to be small, the stored energy can be time-averaged similar to sinusoidal fields. Thus the time average stored energy has the form

$$\langle W_T\rangle = \langle W_E\rangle + \langle W_H\rangle \cong W_0 e^{-2\alpha t} \tag{26.1.7}$$

If there is no loss, $\langle W_T\rangle$ will remain constant. However, with loss, the average stored energy will decrease to $1/e$ of its original value at $t = \tau = \frac{1}{2\alpha}$. The Q of a cavity is defined as the number of free oscillations in radians (rather than cycles) that the field undergoes before the energy stored decreases to $1/e$ of its original value (see Figure 26.2). In a time interval $\tau = \frac{1}{2\alpha}$, the number of free oscillations in radians is $\omega\tau$ or $\frac{\omega}{2\alpha}$; hence, the Q is defined to be [34]

$$Q \cong \frac{\omega}{2\alpha} \tag{26.1.8}$$

[3]Remember from your signals and systems course that poles have to occur in conjugate pair due to the real-value nature of the signal.

Here, Q is an approximate concept, and makes sense only if the system has low loss.

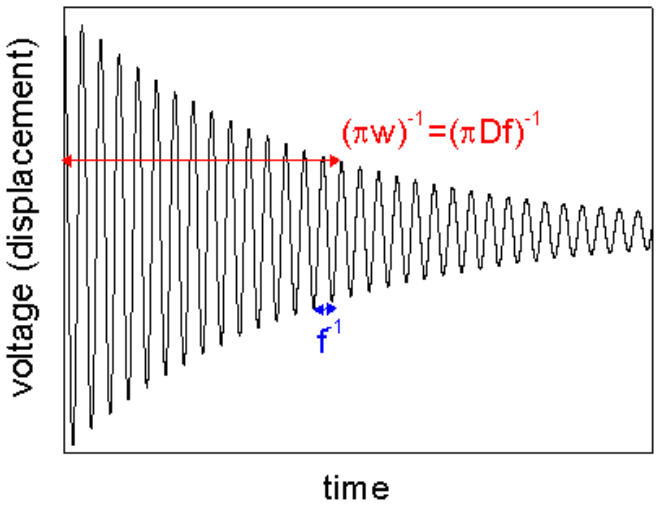


Figure 26.2: A typical time domain response of a high Q system (courtesy of Wikipedia).

The cavity field diminishes with time because of losses in the cavity. These losses can come about because of imperfect conductor on the cavity wall (also called metal loss, or copper loss), materials can have complex permittivity or a loss tangent, or due to radiation if the cavity has a hole. Furthermore, by energy conservation, the decrease in stored energy per unit time must be equal to the total power dissipated in the losses of a cavity. In other words,

$$\langle P_D \rangle = -\frac{d\langle W_T \rangle}{dt} \tag{26.1.9}$$

By further assuming that $\langle W_T \rangle$ has to be of the form in (26.1.7), then

$$-\frac{d\langle W_T \rangle}{dt} \cong 2\alpha W_0 e^{-2\alpha t} = 2\alpha \langle W_t \rangle \tag{26.1.10}$$

From the above, we can estimate the decay constant

$$\alpha \cong \frac{\langle P_D \rangle}{2\langle W_T \rangle} \tag{26.1.11}$$

Hence, we can rewrite equation (26.1.8) for the Q as

$$Q \cong \frac{\omega \langle W_T \rangle}{\langle P_D \rangle} \tag{26.1.12}$$

By further letting $\omega = 2\pi/T$, we lend further physical interpretation to express Q as

$$Q \cong 2\pi \frac{\langle W_T \rangle}{\langle P_D \rangle T} = 2\pi \frac{\text{Total Energy Stored}}{\text{Energy Dissipated/Cycle}} \tag{26.1.13}$$

In the above, Q is a concept defined at the resonant frequency of the resonator, and hence, we evaluate $\omega = \omega_R$.

In a cavity, the energy can dissipate in either the dielectric loss or the wall loss of the cavity due to the finiteness of the conductivity. It has to be re-emphasized that the Q is a low-loss, asymptotic concept, and hence, the above formulas are only approximately true.

26.1.2 Relation to Bandwidth and Pole Location

As seen from above, the resonance of a system is related to the pole of the transfer function. For instance, in our previous example of the RLC tank circuit, the admittance $Y(\omega)$ can be thought of as a transfer function in linear system theory: The input is the voltage $V(\omega)$, while the output is the current $I(\omega)$. If we encounter the resonance of the system at a particular frequency, the transfer function becomes infinite. This infinite value can be modeled by a pole of the transfer function in the complex ω plane.

For a system with a simple pole, the transfer function $Y(\omega)$ in (26.1.4) of the system can be written as

$$Y(\omega) = \frac{A}{j\omega - j\omega_R} + \frac{A^*}{-j\omega + j\omega_R^*} \tag{26.1.14}$$

The second term is added to give this transfer function conjugate symmetry [55] [Table 5.1], a necessity for real-valued signals. In the above, where we have assumed that $\omega_R = \omega_0 + j\alpha$, the resonant frequency is complex. In general, to find the current $i(t)$, we take the inverse Fourier transform of $I(\omega)$, or that

$$i(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(\omega) V(\omega) e^{j\omega t} d\omega \tag{26.1.15}$$

The above does guarantee the real-valuedness of $i(t)$ because of the way we have defined $Y(\omega)$. For the other real-valuedness requirements, we have $V(-\omega) = V^*(\omega)$. The path of integration (also called Fourier inversion contour/path) above is along the real ω axis.[4]. Because of the loss in the system, the poles are never encountered on the real ω axis, and hence, the above Fourier inverse transform is well defined and unique.[5]

In principle, when $\omega = \omega_R$, the transfer function $Y(\omega)$ becomes infinite, but this does not happen in practice because ω_R is complex, and ω, the variable of integration, or the operating frequency is real. In other words, when the pole is displaced slightly off the real axis to account for loss gives rise to a well-defined Fourier inversion path, yielding uniqueness to the solution.

For frequency close to ω_R, we can use a single pole approximation to evaluate the integrand. It is quite clear that $|Y(\omega)|$ would peak at $\omega = \omega_0$. At $\omega = \omega_0 \pm \alpha$, the magnitude of $|Y(\omega)|$ will be $1/\sqrt{2}$ smaller, or that the power, which is proportional to $|Y(\omega)|^2$, will be half as small. Therefore, the half-power points compared to the peak are at $\omega = \omega_0 \pm \alpha$. We can surmise this behavior by studying the magnitude of $|1/(\omega - \omega_R)|$ with respect to Figure 26.3.[6]

[4] If one derives Fourier transform from Fourier series, the frequencies are all real to begin with, and hence, the final Fourier inversion path is on the real axis [55]

[5] Does this ring familiar? We have to introduce loss in the electromagnetic system in order to prove uniqueness of the solution!

[6] In many undergraduate electrical engineering textbooks, Bode plots are introduced to illustrate them.

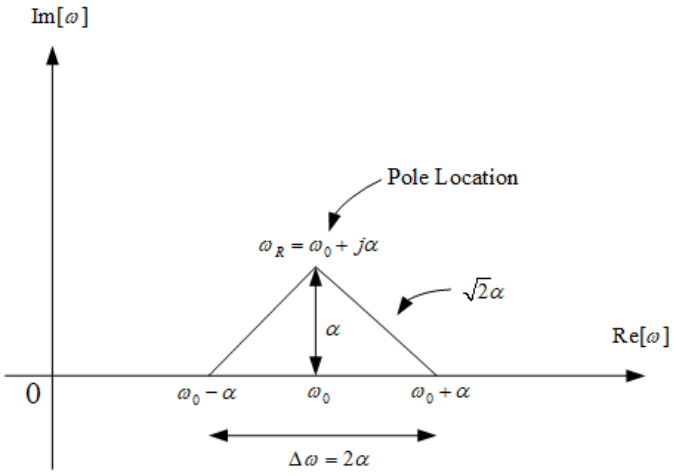


Figure 26.3: From this figure, it is seen that $\omega - \omega_R$ is smallest when $\omega = \omega_0$. Hence, $1/(\omega - \omega_R)$ is largest when $\omega = \omega_0$ on the real axis. Furthermore, its magnitude diminishes to $1/\sqrt{2}$ of its peak value when $\omega = \omega_0 \pm \alpha$. These are the half-power points; hence, the full-width half-maximum (FWHM) bandwidth is $\Delta\omega = \alpha$. This is also the half-power bandwidth.

Thus, the full-width half maximum (FWHM) bandwidth is defined to be $\Delta\omega = 2\alpha$. And the Q can be written as in terms of the half-power bandwidth $\Delta\omega$ of the system, viz.,

$$Q = \omega/(2\alpha) \cong \omega_0/\Delta\omega \tag{26.1.16}$$

Since Q is a narrow-band concept, we can assume that $\omega \approx \omega_0$. Again, we bear in mind that Q is defined at the resonance of the problem and it is a low-loss concept, and hence, $\omega \cong \omega_0$. The above implies that the narrower the bandwidth, the higher the Q of the system. Typical plots of transfer function versus frequency for a system with different Q's are shown in Figure 26.4.

The physics of the Q of an electromagnetic cavity is similar to that of an RLC tank circuit. It is much harder to derive the input-output relation for a general electromagnetic system, other than the use of impedance matrix $\overline{\mathbf{Z}}$ system. In this case, the input/output of the linear system is given by

$$\mathbf{I} = \overline{\mathbf{Y}} \cdot \mathbf{V} \tag{26.1.17}$$

where $\overline{\mathbf{Y}} = (\overline{\mathbf{Z}})^{-1}$, the inverse of the impedance matrix.

It is beyond the scope of this course, but it can be shown that the resonant modes of the system correspond to the zeroes of the determinant of the impedance matrix $\overline{\mathbf{Z}}$ or $f(\omega) = \det(\overline{\mathbf{Z}})$. Here, $f(\omega)$ will exhibit poles as in the RLC tank circuit. In fact, there are some scholars who use circuit theory to model Maxwell's equations called the TLM (transmission-line matrix) method [102].

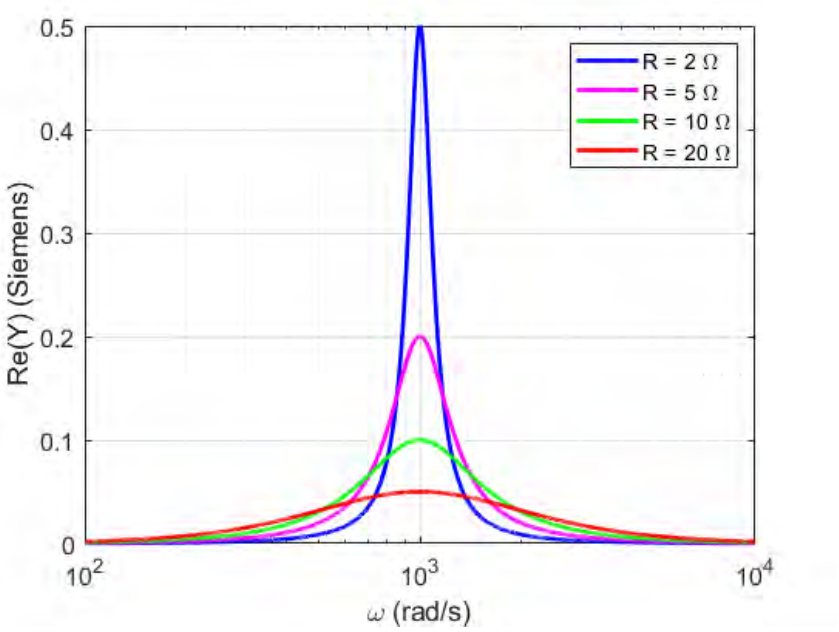


Figure 26.4: A typical system response versus frequency for different Q's using (26.1.4). The Q is altered by changing the resistor R in the circuit. Only the real part of Y is plotted.

26.1.3 Wall Loss and Q for a Metallic Cavity—A Perturbation Concept

To estimate the Q of a cavity, we will need to calculate the loss inside the cavity as well as the energy stored according to (26.1.12). We can use perturbation concept to estimate the Q. First, we assume a lossless cavity so that the cavity wall is made from PEC. In this case, $\hat{n} \times \mathbf{E} = 0$ and no power can be absorbed by the waveguide wall. Then we assume a small loss by assuming now that the cavity wall is made of imperfect conductors and hence, $\hat{n} \times \mathbf{E} \neq 0$ but small. We can assume that the magnetic field $\mathbf{H}$ remains largely unchanged before and after we have introduced loss: for a PEC surface, a nonzero $\hat{n} \times \mathbf{H}$ is needed to support a surface current. But $\hat{n} \times \mathbf{E}$ is zero for a PEC and becomes nonzero for imperfect conductors.

If the cavity is filled with air, then the loss comes mainly from the metallic loss or copper-loss from the cavity wall. In this case, the time-average power dissipated on the wall is given by [34]

$$\langle P_D \rangle = \frac{1}{2} \Re e \oint_S (\mathbf{E} \times \mathbf{H}^*) \cdot \hat{n}\, dS = \frac{1}{2} \Re e \oint_S (\hat{n} \times \mathbf{E}) \cdot \mathbf{H}^*\, dS \tag{26.1.18}$$

where we have used the scalar-triple product identity to rewrite the integrand, and S is the surface of the cavity wall.[7] Here, $(\hat{n} \times \mathbf{E})$ is the tangential component of the electric field which would

[7] We have used the cyclic identity that $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$ in the above (see Some Useful Mathematical Formulas).

have been zero if the cavity wall is made of ideal PEC. Also, $\hat{n}$ is taken to be the outward pointing normal at the surface S. The β (or k) vector in the transmitted medium is very large due to the high-conductivity of the wall. Due to the phase-matching condition, the transmitted wave vector β is almost normal to the interface.[8] Therefore, we can approximate the transmitted wave as a wave propagating normal to the interface. In the metal, it decays predominantly in the direction of propagation which is normal to the surface as well (see Section 8.1).

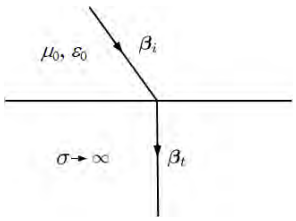


Figure 26.5: Due to high σ in the metal, and large $|\boldsymbol{\beta}_t|$, phase matching condition requires the transmitted wave vector to be almost normal to the surface. Then the field can be assumed to be a normally incident plane wave field in the transmitted or the metal region. The tangential electric field is almost zero, and the tangential magnetic field remains almost the same using this perturbative concept.

For such a wave, we can approximate $\hat{n} \times \mathbf{E} = \mathbf{H}_t Z_m$ where Z_m is the intrinsic impedance for the metallic conductor, as shown in Section 8.1, which is $Z_m = \sqrt{\frac{\mu}{\varepsilon_m}} \approx \sqrt{\frac{\mu}{-j\frac{\sigma}{\omega}}} = \sqrt{\frac{\omega\mu}{2\sigma}}(1+j)$,[9] where we have assumed that $\varepsilon_m \approx -\frac{j\sigma}{\omega}$, and $\mathbf{H}_t$ is the tangential magnetic field. From these equations, we can see that the tangential $\mathbf{E}$ field is small but tangential $\mathbf{H}$ field is not small. It is to be noted that the above approximation is only valid on a flat surface, which is mostly true on a large part of the cavity wall.

This relation between $\mathbf{E}$ and $\mathbf{H}$ will ensure that power is flowing into the metallic surface. Therefore,

$$\langle P_D \rangle = \frac{1}{2}\Re e \oint_S \sqrt{\frac{\omega\mu}{2\sigma}}(1+j)\,|\mathbf{H}_t|^2\,dS = \frac{1}{2}\sqrt{\frac{\omega\mu}{2\sigma}} \oint_S |\mathbf{H}_t|^2\,dS \tag{26.1.19}$$

By further assuming that the stored electric and magnetic energies of a cavity are equal to each other at resonance,[10] the stored energy can be obtained by

$$\langle W_T \rangle = \frac{1}{2}\mu \int_V |\mathbf{H}|^2 dV = \frac{1}{2}\varepsilon \int_V |\mathbf{E}|^2 dV \tag{26.1.20}$$

[8] The horizontal components of the wave vectors $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_t$ have to be real for phase matching. Hence, the $\boldsymbol{\beta}_t$ is only complex in the normal direction.

[9] When an electromagnetic wave enters a conductive region with a large β, it can be shown that the wave is refracted to propagate normally to the surface as shown in Figure 26.5, and hence, this formula can be applied. Moreover, the Fresnel reflection and transmission coefficients still apply here.

[10] They are definitely equal to each other for the lossless case. For the lossy case, they are approximately equal.

Written explicitly, the $Q \cong \frac{\omega\langle W_T\rangle}{\langle P_D\rangle}$ becomes

$$Q = \sqrt{2\omega\mu\sigma}\frac{\int_V |\mathbf{H}|^2\, dV}{\oint_S |\mathbf{H}_t|^2\, dS} = \frac{2}{\delta}\frac{\int_V |\mathbf{H}|^2\, dV}{\oint_S |\mathbf{H}_t|^2\, dS} \tag{26.1.21}$$

In the above, δ is the skin depth of the metallic wall. Hence, the more energy stored there is with respect to the power dissipated, the higher the Q of a resonating system. Also, the lower the metal loss, or the smaller the skin depth, the higher the Q would be.

Notice that in (26.1.21), the numerator is a volume integral and hence, is proportional to volume, while the denominator is a surface integral and is proportional to surface. Thus, the Q, a dimensionless quantity, is roughly proportional to

$$Q \sim \frac{V}{S\delta}A \tag{26.1.22}$$

where V is the volume of the cavity, while S is its surface area, and A is a dimensionless constant yet to be determined. From the above, it is noted that a large cavity compared to its skin depth has a larger Q than a small cavity. Thus in microwave engineering, a large cavity (compared to wavelength) is needed to obtain a high Q resonance.

It is easy to make large cavities in optics as the wavelength is small. Also, dielectric resonator cavity can be made out of glass where the primary loss will be from the material. Quality factor $\approx 10^8 \sim 10^9$ is possible [180]. A dielectric resonator using total internal reflection to trap the wave as it bounces around is called a whispering gallery mode resonator. The glass can be made with very low loss.[11] This together with the large size of the cavity compared to wavelength gives the resonator very high Q. The large cavity size increases the stored energy as well, which is good for increasing its Q as evident from (26.1.22).

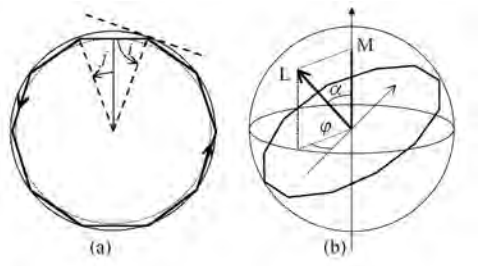


Figure 26.6: A dielectric resonator can be made by using total internal reflection of the bouncing wave within the resonator [180]. Such a mode is called a whispering gallery mode. It has high Q because the glass making the cavity has little loss, and that the cavity can be made very large compared to wavelength, increasing the stored energy.

[11] Silica is a gift of God. With it, we have the optical fiber with little loss, and the semiconductor devices which have revolutionized our modern world.

26.1.4 Example: The Q of TM_{110} Mode

For the TM_{110} mode, as can be seen from the previous lecture, the only electric field is $\mathbf{E} = \hat{z}E_z$, with $\partial/\partial z = 0$. Then with $m = 1$, $n = 1$, and $p = 0$, we have

$$E_z = E_0 \sin\left(\frac{\pi x}{a}\right) \sin\left(\frac{\pi y}{b}\right) \tag{26.1.23}$$

The magnetic field can be derived from the electric field using Maxwell's equation or Faraday's law, giving

$$H_x = \frac{j\omega\epsilon}{\omega^2\mu\epsilon}\frac{\partial}{\partial y}E_z = \frac{j\left(\frac{\pi}{b}\right)}{\omega\mu}E_0 \sin\left(\frac{\pi x}{a}\right)\cos\left(\frac{\pi y}{b}\right) \tag{26.1.24}$$

$$H_y = \frac{-j\omega\epsilon}{\omega^2\mu\epsilon}\frac{\partial}{\partial x}E_z = -\frac{j\left(\frac{\pi}{a}\right)}{\omega\mu}E_0 \cos\left(\frac{\pi x}{a}\right)\sin\left(\frac{\pi y}{b}\right) \tag{26.1.25}$$

Therefore[12]

$$\begin{aligned}
\int_V |\mathbf{H}|^2\, dV &= \int_{-d}^{0}\int_0^b\int_0^a dx\, dy\, dz \left[|H_x|^2 + |H_y|^2\right] \\
&= \frac{|E_0|^2}{\omega^2\mu^2}\int_{-d}^{0}\int_0^b\int_0^a dx\, dy\, dz \\
&\quad \left[\left(\frac{\pi}{b}\right)^2 \sin^2\left(\frac{\pi x}{a}\right)\cos^2\left(\frac{\pi y}{b}\right) + \left(\frac{\pi}{a}\right)^2 \cos^2\left(\frac{\pi x}{a}\right)\sin^2\left(\frac{\pi y}{b}\right)\right] \\
&= \frac{|E_0|^2}{\omega^2\mu^2}\frac{\pi^2}{4}\left[\frac{a}{b} + \frac{b}{a}\right] d
\end{aligned} \tag{26.1.26}$$

A cavity has six faces, finding the tangential exponent at each face and integrate

$$\begin{aligned}
\oint_S |\mathbf{H}_t|\, dS &= 2\int_0^b\int_0^a dx\, dy \left[|H_x|^2 + |H_y|^2\right] \\
&\quad + 2\int_{-d}^{0}\int_0^a dx\, dz\, |H_x(y=0)|^2 + 2\int_{-d}^{0}\int_0^b dy\, dz\, |H_y(x=0)|^2 \\
&= \frac{2|E_0|^2}{\omega^2\mu^2}\frac{\pi^2 ab}{4}\left[\frac{1}{a^2} + \frac{1}{b^2}\right] + \frac{2\left(\frac{\pi}{b}\right)^2}{\omega^2\mu^2}|E_0|^2\frac{ad}{2} + \frac{2\left(\frac{\pi}{a}\right)^2}{\omega^2\mu^2}|E_0|^2\frac{bd}{2} \\
&= \frac{\pi^2|E_0|^2}{\omega^2\mu^2}\left[\frac{b}{2a} + \frac{a}{2b} + \frac{ad}{b^2} + \frac{bd}{a^2}\right]
\end{aligned} \tag{26.1.27}$$

Hence the Q from (26.1.21) is finally given as

$$Q = \frac{1}{2\delta}\frac{\left(\frac{ad}{b} + \frac{bd}{a}\right)}{\left(\frac{b}{2a} + \frac{a}{2b} + \frac{ad}{b^2} + \frac{bd}{a^2}\right)} \tag{26.1.28}$$

[12]Since the electric field is simpler than the magnetic field, it is easier to find the energy stored using the electric field. Like an LC tank circuit, the magnetic field energy stored and the electric field energy stored are equal to each other.

The result shows that the larger the cavity, the higher the Q. This is because the Q, as mentioned before, is the ratio of the energy stored in a volume to the energy dissipated over the surface of the cavity.

26.2 Mode Orthogonality and Matrix Eigenvalue Problem

It turns out that the modes of a waveguide or a resonator are orthogonal to each other. This is intimately related to the orthogonality of eigenvectors of a matrix operator.[13] Thus, it is best to understand this by the homomorphism between the electromagnetic mode problem and the matrix eigenvalue problem. Because of this similarity, electromagnetic modes are also called eigenmodes. Thus it is prudent that we revisit the matrix eigenvalue problem (EVP) here.

26.2.1 Hermiticity of an Operator

We can use inner product to define the Hermiticity and the symmetry of an operator. This is great because it can be easily extended to infinite dimensional Hilbert spaces.

Transposes of a Matrix Operator

The transpose of a matrix operator is defined, with matrix operator notation, to be such that[14]

$$\mathbf{f}^t \cdot \overline{\mathbf{G}} \cdot \mathbf{g} = \mathbf{g}^t \cdot \overline{\mathbf{G}}^t \cdot \mathbf{f} \tag{26.2.1}$$

If $\overline{\mathbf{G}}^t = \overline{\mathbf{G}}$, then $\overline{\mathbf{G}}$ is symmetric. In other words, its transpose is itself.

The conjugate (Hermitian) transpose of a matrix operator is defined, to be such that

$$\mathbf{f}^\dagger \cdot \overline{\mathbf{G}} \cdot \mathbf{g} = \mathbf{g}^\dagger \cdot \overline{\mathbf{G}}^\dagger \cdot \mathbf{f} \tag{26.2.2}$$

If $\overline{\mathbf{G}}^\dagger = \overline{\mathbf{G}}$, then $\overline{\mathbf{G}}$ is Hermitian. In other words, its conjugate transpose is itself. These operators are also called self-adjoint operators.

The above definition is convenient because it can be extended to infinite dimensional spaces which are usually called Hilbert spaces. Then the inner product between two vectors becomes functional inner product.

$$\mathbf{f}^t \cdot \mathbf{g} \to \langle f, g \rangle \tag{26.2.3}$$

where the inner product on the right-hand side is defined as

$$\langle f, g \rangle = \int f(\mathbf{r}) g(\mathbf{r}) d\mathbf{r} \tag{26.2.4}$$

The above inner product exists if the integral converges.

[13] This mathematical "homomorphism" is not discussed in any other electromagnetic textbooks.

[14] The transpose of a scalar is a scalar, and then we use the rule for the transpose of product of matrix operators [181].

26.2.2 Matrix Eigenvalue Problem (EVP)

It is known in matrix theory that if a matrix is Hermitian, then its eigenvalues are all real. Furthermore, their eigenvectors with distinct eigenvalues are orthogonal to each other [83]. Assume that an eigenvalue and an eigenvector exists for the Hermitian matrix $\overline{\mathbf{A}}$. Then

$$\overline{\mathbf{A}} \cdot \mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{26.2.5}$$

To prove the real value of λ_i, we dot multiply the above from the left by $\mathbf{v}_i^\dagger$ where $\dagger$ indicates conjugate transpose. Then the above becomes

$$\mathbf{v}_i^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{v}_i = \lambda_i \mathbf{v}_i^\dagger \cdot \mathbf{v}_i \tag{26.2.6}$$

Since $\overline{\mathbf{A}}$ is Hermitian, or $\overline{\mathbf{A}}^\dagger = \overline{\mathbf{A}}$, then the quantity $\mathbf{v}_i^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{v}_i$ is purely real. Moreover, the quantity $\mathbf{v}_i^\dagger \cdot \mathbf{v}_i$ is positive real.[15] So in order for the above to be satisfied, λ_i has to be real.

To prove orthogonality of eigenvectors, now, assume that $\overline{\mathbf{A}}$ has two eigenvectors with distinct eigenvalues such that

$$\overline{\mathbf{A}} \cdot \mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{26.2.7}$$

$$\overline{\mathbf{A}} \cdot \mathbf{v}_j = \lambda_j \mathbf{v}_j \tag{26.2.8}$$

Left dot multiply the first equation with $\mathbf{v}_j^\dagger$ and do the same to the second equation with $\mathbf{v}_i^\dagger$, one gets

$$\mathbf{v}_j^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{v}_i = \lambda_i \mathbf{v}_j^\dagger \cdot \mathbf{v}_i \tag{26.2.9}$$

$$\mathbf{v}_i^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{v}_j = \lambda_j \mathbf{v}_i^\dagger \cdot \mathbf{v}_j \tag{26.2.10}$$

Taking the conjugate transpose of (26.2.9) in the above, and since $\overline{\mathbf{A}}$ is Hermitian, their left-hand sides (26.2.9) and (26.2.10) are the same. Subtracting the two equations, we arrive at

$$0 = (\lambda_i - \lambda_j) \mathbf{v}_j^\dagger \cdot \mathbf{v}_i \tag{26.2.11}$$

For distinct eigenvalues, $\lambda_i \neq \lambda_j$, the only way for the above to be satisfied is that

$$\mathbf{v}_j^\dagger \cdot \mathbf{v}_i = C_i \delta_{ij} \tag{26.2.12}$$

Hence, eigenvectors of a Hermitian matrix with distinct eigenvalues are orthogonal to each other. The eigenvalues are also real.

26.2.3 Power Orthogonality

The subject of power orthogonality is intimately related to mode orthogonality [101]. Because of mode orthogonality, one can show that each individual mode in a waveguide carries power separately from each other, which is power orthogonality. This power orthogonality can be generalized to plane-wave modes in free space.

[15] Convince yourself that this is the case if you are dubious.

Given an incident plane wave impinging on a scatterer, only the plane wave with plane-wave **k** vector that is parallel to the **k** vector of the incident plane wave can carry energy away from the incident plane wave due to power orthogonality. Hence, the total power scattered by a scatterer with plane wave impinging on it is proportional to the power scattered in the forward direction. This is the statement of the optical theorem [49].

Exercises for Lecture 26

Problem 26-1: For a rectangular metallic cavity, show that the cavity modes can also derived using a scalar potential approach.

(i) Show that for:

$$\text{TM modes,} \quad \mathbf{H}(\mathbf{r}) = \nabla \times \hat{z}\Psi_e(\mathbf{r})$$

$$\text{TE modes,} \quad \mathbf{E}(\mathbf{r}) = \nabla \times \hat{z}\Psi_h(\mathbf{r})$$

(ii) Find the scalar wave equations that $\Psi_e(\mathbf{r})$ and $\Psi_h(\mathbf{r})$ satisfy.

(iii) Find the boundary conditions on the six walls of the cavity for $\Psi_e(\mathbf{r})$ and $\Psi_h(\mathbf{r})$.

(iv) Show that the modes of a rectangular cavity are orthogonal to each other (but this is not restricted to rectangular cavities only). Show that the scalar wave modes found in (ii) and (iii) are orthogonal to each other. Namely, that

$$\int_V dV \Psi_{mnl}(\mathbf{r})\Psi_{m'n'l'}(\mathbf{r}) = C\delta_{mm'}\delta_{nn'}\delta_{ll'}$$

where Ψ above can be either Ψ_e or Ψ_h. Show that because of this, the fields are also orthogonal, namely that

$$\int_V dV \mathbf{H}_{mnl}(\mathbf{r}) \cdot \mathbf{H}_{m'n'l'}(\mathbf{r}) = C'\delta_{mm'}\delta_{nn'}\delta_{ll'}$$

The above volume integrals are over the volume of the cavity.

Part III

Radiation, High-Frequency Approximation, Computational Electromagnetics, Quantum Theory of Light

Chapter 27

Scalar and Vector Potentials

27.1 Scalar and Vector Potentials for Time-Harmonic Fields

Now that we have studied the guidance of waves by waveguides, and the trapping of electromagnetic waves by cavity resonators, it will be interesting to consider how electromagnetic waves radiate from sources. This is best done via the scalar and vector potential formulation.

Previously, we have studied the use of scalar potential Φ for electrostatic problems. Then we learnt the use of vector potential $\mathbf{A}$ for magnetostatic problems. Now, we will study the combined use of both scalar and vector potentials concurrently for solving time-harmonic (electrodynamic) problems.

This is important for bridging the gap between the static regime where the frequency is zero or low, and the dynamic regime where the frequency is not low. For the dynamic regime, it is important to understand the radiation of electromagnetic fields which has a plethora of advanced applications. Electrodynamic regime is important for studying antennas, communications, sensing, wireless power transfer applications, and many more. High frequency electromagnetics is important in THz technologies, optics, nano-lithography, and quantum technologies. Hence, it is imperative that we understand how time-varying electromagnetic fields radiate from sources.

It is also crucial to understand when static or circuit (quasi-static) regimes are important. The circuit theory has been used to solve many highly complex problems. These solutions have fueled the microchip, integrated circuit design (ICD) industry, millimeter wave integrated circuits, as well as the antenna design industry. There is also the emerging area of antenna on a chip (AoC) and antenna in a package (AiP) [182, 183]. Therefore, it is beneficial to understand when electromagnetic problems can be approximated with simple circuit problems and solved using simple laws such as Kirchhoff current law (KCL) an Kirchhoff voltage law (KVL).

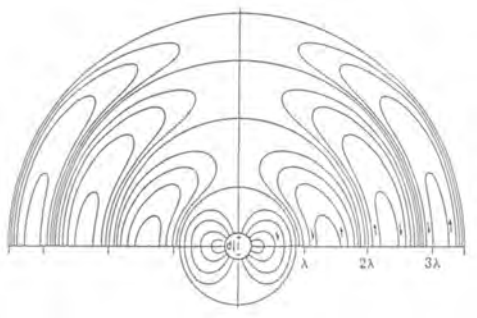


Figure 27.1: Plot of a snapshot of the time-harmonic electric field around a dipole source which is time-varying, where only the field in the upper half-space is shown. Close to the source, the field resembles that of a static electric dipole, but far away from the source, the electromagnetic field is detached from the source: due to the finite velocity of light, the electromagnetic field cannot keep up with the changing dipole field near the source. In other words, the source starts to shed energy to the far region and radiate.

27.2 Scalar and Vector Potentials for Statics—A Review

Previously, we have studied scalar and vector potentials for electrostatics and magnetostatics where the frequency ω is identically zero. The four Maxwell's equations for a homogeneous media are then

$$\nabla \times \mathbf{E} = 0 \tag{27.2.1}$$

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{27.2.2}$$

$$\nabla \cdot \varepsilon \mathbf{E} = \varrho \tag{27.2.3}$$

$$\nabla \cdot \mu \mathbf{H} = 0 \tag{27.2.4}$$

Looking at the first equation above, and using the knowledge that $\nabla \times (\nabla \Phi) = 0$, we can construct a solution to (27.2.1) easily. Thus, in order to satisfy the first of Maxwell's equations or Faraday's law above, we let

$$\mathbf{E} = -\nabla \Phi \tag{27.2.5}$$

The above implies that (27.2.1) is satisfied. It also implies that the static field is conservative, or that $\oint \mathbf{E} \cdot d\mathbf{l} = 0$, implying that no net work is done if one moves a charge in a closed loop. Using the (27.2.5) in (27.2.3), we get,

$$\nabla \cdot \varepsilon \nabla \Phi = -\varrho \tag{27.2.6}$$

Then for a homogeneous medium where ε is a constant, $\nabla \cdot \varepsilon \nabla \Phi = \varepsilon \nabla \cdot \nabla \Phi = \varepsilon \nabla^2 \Phi$, and we have

$$\nabla^2 \Phi = -\frac{\varrho}{\varepsilon} \tag{27.2.7}$$

which is the Poisson's equation for electrostatics for a homogeneous medium.

Now looking at (27.2.4) where $\nabla \cdot \mu\mathbf{H} = 0$, we let

$$\mu\mathbf{H} = \nabla \times \mathbf{A} \tag{27.2.8}$$

Since $\nabla \cdot (\nabla \times \mathbf{A}) = 0$, the last of Maxwell's equations (27.2.4) is automatically satisfied. Using the above in the second of Maxwell's equations above, we get

$$\nabla \times \nabla \times \mathbf{A} = \mu\mathbf{J} \tag{27.2.9}$$

Now, using the fact that $\nabla \times \nabla \times \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2\mathbf{A}$, and Coulomb gauge that $\nabla \cdot \mathbf{A} = 0$, we arrive at

$$\nabla^2\mathbf{A} = -\mu\mathbf{J} \tag{27.2.10}$$

which is the vector Poisson's equation for a homogeneous medium. We will repeat the above derivation when $\omega \neq 0$.

27.2.1 Scalar and Vector Potentials for Electrodynamics

Since dynamic or time-varying problems are of utmost importance in electromagnetics, we will study it next. To this end, assuming linearity, we will start with frequency domain Maxwell's equations with sources $\mathbf{J}$ and ϱ included, and later, see how these sources $\mathbf{J}$ and ϱ can radiate electromagnetic fields. Maxwell's equations in the frequency domain are[1]

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{27.2.11}$$
$$\nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E} + \mathbf{J} \tag{27.2.12}$$
$$\nabla \cdot \varepsilon\mathbf{E} = \varrho \tag{27.2.13}$$
$$\nabla \cdot \mu\mathbf{H} = 0 \tag{27.2.14}$$

We can view the above as a set of linear system of equations where $\mathbf{J}$ and ϱ are inputs, and $\mathbf{E}$ and $\mathbf{H}$ are outputs.

In order to satisfy the last Maxwell's equation, as before, we let

$$\mu\mathbf{H} = \nabla \times \mathbf{A} \tag{27.2.15}$$

Now, using (27.2.15) in (27.2.11), we have

$$\nabla \times (\mathbf{E} + j\omega\mathbf{A}) = 0 \tag{27.2.16}$$

Since $\nabla \times (\nabla\Phi) = 0$, the above implies that $\mathbf{E} + j\omega\mathbf{A} = -\nabla\Phi$, or that

$$\mathbf{E} = -j\omega\mathbf{A} - \nabla\Phi \tag{27.2.17}$$

The above indicates that the electrostatic theory of letting $\mathbf{E} = -\nabla\Phi$ we have learnt previously in Section 3.3.1 is not exactly correct when $\omega \neq 0$. The $-j\omega\mathbf{A}$ term above, in accordance to Faraday's

[1]The time domain field is a linear superposition of many frequency domain fields. We can get to frequency domain Maxwell's equations by either using phasor technique or Fourier transform technique.

law, is the contribution to the electric field from the time-varying magnetic field, and hence, is termed the induction term.[2]

Furthermore, the above shows that once $\mathbf{A}$ and Φ are known, one can determine the fields $\mathbf{H}$ and $\mathbf{E}$ assuming that $\mathbf{J}$ and ϱ are given. To this end, we will derive equations for $\mathbf{A}$ and Φ in terms of the sources $\mathbf{J}$ and ϱ. Substituting (27.2.15) and (27.2.17) into (27.2.12) gives

$$\nabla \times (\nabla \times \mathbf{A}) = j\omega\mu\varepsilon(-j\omega\mathbf{A} - \nabla\Phi) + \mu\mathbf{J} \tag{27.2.18}$$

Or upon rearrangement, after using that $\nabla \times (\nabla \times \mathbf{A}) = \nabla\nabla \cdot \mathbf{A} - \nabla \cdot \nabla\mathbf{A}$, we have

$$\nabla^2\mathbf{A} + \omega^2\mu\varepsilon\mathbf{A} = -\mu\mathbf{J} + j\omega\mu\varepsilon\nabla\Phi + \nabla\nabla \cdot \mathbf{A} \tag{27.2.19}$$

Moreover, using (27.2.17) in (27.2.13), we have

$$\nabla \cdot (j\omega\mathbf{A} + \nabla\Phi) = -\frac{\varrho}{\varepsilon} \tag{27.2.20}$$

In the above, (27.2.19) and (27.2.20) represent two equations for the two unknowns $\mathbf{A}$ and Φ, expressed in terms of the known quantities, the sources $\mathbf{J}$ and ϱ. But these equations are coupled to each other. They look complicated and are rather unwieldy to solve at this point.

Fortunately, the above can be simplified! As in the magnetostatic case, the vector potential $\mathbf{A}$ in (27.2.15) is not unique. To show this, one can always construct a new $\mathbf{A}' = \mathbf{A} + \nabla\Psi$ that produces the same magnetic field $\mu\mathbf{H}$ via (27.2.15), since $\nabla \times (\nabla\Psi) = 0$. It is quite clear that $\mu\mathbf{H} = \nabla \times \mathbf{A} = \nabla \times \mathbf{A}'$. Moreover, one can further show that Φ is also non-unique [49]. Namely, with

$$\mathbf{A}' = \mathbf{A} + \nabla\Psi \tag{27.2.21}$$

$$\Phi' = \Phi - j\omega\Psi \tag{27.2.22}$$

it can be shown that the new $\mathbf{A}'$ and Φ' produce the same $\mathbf{E}$ and $\mathbf{H}$ field. The above is known as gauge transformation [49], clearly showing the non-uniqueness of $\mathbf{A}$ and Φ.

To make them unique, in addition to specifying what $\nabla \times \mathbf{A}$ should be in (27.2.15), we need to specify its divergence or $\nabla \cdot \mathbf{A}$ as in the electrostatic case.[3] A clever way to specify the divergence of $\mathbf{A}$ is to choose it to simplify the complicated equations above in (27.2.19). We choose a gauge so that the last two terms on the right-hand side in equation (27.2.19) cancel each other. In other words, we let

$$\nabla \cdot \mathbf{A} = -j\omega\mu\varepsilon\Phi \tag{27.2.23}$$

The above is judiciously chosen so that the pertinent equations (27.2.19) and (27.2.20) will be simplified and decoupled. With the use of (27.2.23) in (27.2.19) and (27.2.20), they now become

$$\nabla^2\mathbf{A} + \omega^2\mu\varepsilon\mathbf{A} = -\mu\mathbf{J} \tag{27.2.24}$$

$$\nabla^2\Phi + \omega^2\mu\varepsilon\Phi = -\frac{\varrho}{\varepsilon} \tag{27.2.25}$$

[2]Notice that in electrical engineering, most concepts related to magnetic fields are inductive!

[3]This is akin to that given a vector $\mathbf{A}$, and an arbitrary vector $\mathbf{k}$, in addition to specifying what $\mathbf{k} \times \mathbf{A}$ is, it is also necessary to specify what $\mathbf{k} \cdot \mathbf{A}$ is to uniquely specify $\mathbf{A}$.

Equation (27.2.23) is known as the Lorenz gauge[4] and the above equations are Helmholtz equations with source terms. These are the equations from which we can solve for the scalar potential Φ and the vector potential $\mathbf{A}$ given the sources $\mathbf{J}$ and ϱ. Not only are these equations simplified, they can be solved independently of each other since they are decoupled from each other.

Equations (27.2.24) and (27.2.25) can be solved using the Green's function method we have learnt previously. Equation (27.2.24) in cartesian coordinates (only in cartesian coordinates) actually constitutes three scalar equations for the three x, y, z components, namely that

$$\nabla^2 A_i + \omega^2 \mu\varepsilon A_i = -\mu J_i \tag{27.2.26}$$

where i above can be x, y, or z. Therefore, (27.2.24) and (27.2.25) together constitute four scalar equations similar to each other. Hence, we need only to solve their point-source response, or the Green's function of these equations by solving

$$\nabla^2 g(\mathbf{r}, \mathbf{r}') + \beta^2 g(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}') \tag{27.2.27}$$

where $\beta^2 = \omega^2 \mu\varepsilon$. This Green's function is sometimes referred to as the fundamental solution or the canonical solution of the problem.

Previously, we have shown that when $\beta = 0$,

$$g(\mathbf{r}, \mathbf{r}') = g(|\mathbf{r} - \mathbf{r}'|) = \frac{1}{4\pi|\mathbf{r} - \mathbf{r}'|}$$

When $\beta \neq 0$, the correct solution is

$$g(\mathbf{r}, \mathbf{r}') = g(|\mathbf{r} - \mathbf{r}'|) = \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r} - \mathbf{r}'|} \tag{27.2.28}$$

which can be verified by back substitution or derived [34] [108, p. 26].

By using the principle of linear superposition, or convolution, the solutions to (27.2.24) and (27.2.25) are then

$$\mathbf{A}(\mathbf{r}) = \mu \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') g(|\mathbf{r} - \mathbf{r}'|) = \mu \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r} - \mathbf{r}'|} \tag{27.2.29}$$

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon} \iiint_V d\mathbf{r}' \varrho(\mathbf{r}') g(|\mathbf{r} - \mathbf{r}'|) = \frac{1}{\varepsilon} \iiint_V d\mathbf{r}' \varrho(\mathbf{r}') \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r} - \mathbf{r}'|} \tag{27.2.30}$$

In the above $d\mathbf{r}'$ is the shorthand notation for $dx'\, dy'\, dz'$: they are volume integrals. The above are three-dimensional convolutional integrals in space.

27.2.2 Degree of Freedom in Maxwell's Equations

From (27.2.24) and (27.2.25), if $\mathbf{J}$ and ϱ are independent, these two equations are independent. Howevever, $\mathbf{J}$ and ϱ are related by the current continuity equation, $\nabla \cdot \mathbf{J} = -j\omega\varrho$. Therefore, using the Lorenz gauge (27.2.23), (27.2.25) can be derived from (27.2.24). Hence, solving (27.2.24)

[4] Please note that this Lorenz is not the same as Lorentz.

suffices to solve Maxwell's equations. Therefore, only three degrees of freedom are needed to solve Maxwell's equations fully for a homogeneous linear medium. Once $\mathbf{A}$ is found, Φ can be found via the Lorenz gauge. But at statics with $\omega = 0$, Φ cannot be derived from $\mathbf{A}$ via the Lorenz gauge. In this case, it appears that four degrees of freedom are needed to fully solve Maxwell's equations. But the following argument will show that still three degrees of freedom are needed.

For the static case, since $\nabla \cdot \mathbf{J} = 0$ via (27.2.2), there are only two degrees of freedom in (27.2.10). This together with (27.2.7), only three degrees of freedom are needed to solve Maxwell's equations as in the dynamic case. It is to be noted that when the scalar and vector potential approach is used to solve Maxwell's equations, all the four Maxwell's equations, (27.2.11) to (27.2.14) are involved. When numerical scheme is developed to solve Maxwell's equations using this approach, the solutions are stable down to very low frequencies [184].

27.2.3 More on Scalar and Vector Potentials

It is to be noted that Maxwell's equations are symmetrical and this is especially so when we add a magnetic current $\mathbf{M}$ to Maxwell's equations and magnetic charge ϱ_m to Gauss's law.[5] Thus the equations become

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} - \mathbf{M} \tag{27.2.31}$$

$$\nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E} + \mathbf{J} \tag{27.2.32}$$

$$\nabla \cdot \mu\mathbf{H} = \varrho_m \tag{27.2.33}$$

$$\nabla \cdot \varepsilon\mathbf{E} = \varrho \tag{27.2.34}$$

The above can be solved in two stages, using the principle of linear superposition because the above is a linear time invariant system. Thus, the sources of the system can be turned on and off consecutively to obtain different solutions to the system. First, we can set $\mathbf{M} = 0$, $\varrho_m = 0$, and $\mathbf{J} \neq 0$, $\varrho \neq 0$, and solve for the fields as we have done before. Second, we can set $\mathbf{J} = 0$, $\varrho = 0$ but let $\mathbf{M} \neq 0$, $\varrho_m \neq 0$ and solve for the fields next. Then the total general solution, by linearity, is just the linear superposition of these two solutions.

For the second case, we set $\mathbf{J} = 0$, $\varrho = 0$ and $\mathbf{M} \neq 0$, $\varrho_m \neq 0$. Then, we can define an electric vector potential $\mathbf{F}$ such that [51, 41]

$$\mathbf{D} = -\nabla \times \mathbf{F} \tag{27.2.35}$$

and a magnetic scalar potential Φ_m such that

$$\mathbf{H} = -\nabla\Phi_m - j\omega\mathbf{F} \tag{27.2.36}$$

By invoking duality principle (see Section 17.2), one gather that [51]

$$\mathbf{F}(\mathbf{r}) = \varepsilon \iiint d\mathbf{r}'\mathbf{M}(\mathbf{r}')g(|\mathbf{r}-\mathbf{r}'|) = \varepsilon \iiint d\mathbf{r}'\mathbf{M}(\mathbf{r}')\frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{27.2.37}$$

$$\Phi_m(\mathbf{r}) = \frac{1}{\mu}\iiint d\mathbf{r}'\varrho_m(\mathbf{r}')g(|\mathbf{r}-\mathbf{r}'|) = \frac{1}{\mu}\iiint d\mathbf{r}'\varrho_m(\mathbf{r}')\frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{27.2.38}$$

[5] In fact, Maxwell himself exploited this symmetry [41].

As mentioned before, even though magnetic sources do not exist, they can be engineered. For instance, an electric current loop antenna resembles a magnetic dipole. Another example is the toroidal antenna shown in Figure 17.12 which resembles a magnetic current loop antenna. In many engineering designs, one can use fictitious magnetic sources to enrich the diversity of electromagnetic technologies.

27.3 When is Static Electromagnetic Theory Valid?

Static electromagnetic theory is often simpler than dynamic electromagnetic theory. It will be prudent to know when we can apply static theory instead of the more complicated electrodynamic theory.

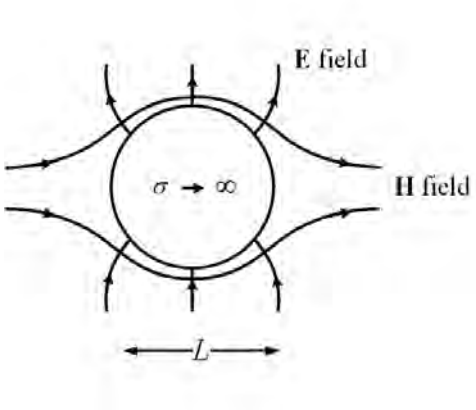


Figure 27.2: The electric and magnetic fields are great contortionists around a perfectly conducting particle. They deform themselves to satisfy the boundary conditions, $\hat{n} \times \mathbf{E} = 0$, and $\hat{n} \cdot \mathbf{H} = 0$ on the PEC surface, even when the particle is very small. In other words, the fields vary on the length-scale of L. Approximately, $\nabla \sim 1/L$ which is large when L is small.

27.3.1 Cutting Through The Chaste

To see when static electromagnetics can be used to approximate electrodynamics, we stare at the Helmholtz equations previously derived, and ask when they can be replaced by Poisson/Laplace equations. They are reproduced here as

$$\nabla^2 \mathbf{A} + \omega^2 \mu \varepsilon \mathbf{A} = -\mu \mathbf{J} \tag{27.3.1}$$

$$\nabla^2 \Phi + \omega^2 \mu \varepsilon \Phi = -\frac{\varrho}{\varepsilon} \tag{27.3.2}$$

By looking at Figure 27.2, in order to satisfy the boundary conditions on the wall of an object, the electromagnetic fields have to contort themselves around the object in order to satisfy the

requisite boundary conditions (as contortionists do in a circus). In order for this to happen, by dimensional analysis[6] $\partial/\partial x$, $\partial/\partial y$, and $\partial/\partial z$ are of the order of $1/L$, or that ∇^2 is of the order of $1/L^2$. Now, we compare the Laplacian operator term, which is of order $1/L^2$, and the $\beta^2 = \omega^2\mu\varepsilon$ term. If $\beta^2 L^2 \ll 1$, then Helmholtz equation is dominated by the Laplacian operator, and it can be replaced by Laplace equations. In this case, static electromagnetics applies.

In the above, $\beta = 2\pi/\lambda$ where λ is the wavelength. Thus, if $L/\lambda \ll 1$, static theory applies. Therefore, in electromagnetics, the yardstick is the wavelength. When the object size is much smaller than the wavelength, we are in the static or quasistatic regime, whereas if the object size is about the wavelength or larger, we are in the electrodynamic regime, or wave-physics regime [185].

27.3.2 An Example

In a word, one can solve, even in optics, where ω is humongous or the wavelength very short, using static analysis if the size of the object L is much smaller than the optical wavelength which is about 400 nm for blue light. Nowadays, plasmonic nano-particles of about 10 nm can be made. If the particle is small enough compared to wavelength of the light, electrostatic analysis can be used to study their interaction with light. And hence, static electromagnetic theory can be used to analyze the wave-particle interaction. This was done in one of the homeworks. This kind of scattering is also referred to as Rayleigh scattering [186].

Figure 27.3 shows an incident light whose wavelength is much longer than the size of the particle. The incident field induces an electric dipole moment on the particle, whose scattered field can be written as

$$\mathbf{E}_s = (\hat{r}2\cos\theta + \hat{\theta}\sin\theta)\left(\frac{a}{r}\right)^3 E_s \tag{27.3.3}$$

while the incident field $\mathbf{E}_0$ and the interior field $\mathbf{E}_i$ to the particle can be expressed as

$$\mathbf{E}_0 = \hat{z}E_0 = (\hat{r}\cos\theta - \hat{\theta}\sin\theta)E_0 \tag{27.3.4}$$

$$\mathbf{E}_i = \hat{z}E_i = (\hat{r}\cos\theta - \hat{\theta}\sin\theta)E_i \tag{27.3.5}$$

By matching boundary conditions, as was done in the homework, it can be shown that

$$E_s = \frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon}E_0 \tag{27.3.6}$$

$$E_i = \frac{3\varepsilon}{\varepsilon_s + 2\varepsilon}E_0 \tag{27.3.7}$$

For a plasmonic nano-particle, for some materials, the particle medium behaves like a plasma (see Chapter 8), and ε_s in the above can be approximately negative, making the denominators of the above expression very close to zero. This is the hallmark of a resonance phenomenon as we have seen in the surface plasmonic polariton, the transverse resonance condition, and the LC tank circuit, and now, the plasmonic nanoparticle. Therefore, the amplitude of the internal and scattered fields can be very large when this happens, and the nano-particles will glitter in the presence of light. Even the ancient Romans realized this! (See Section 8.3.7.)

[6] Which is often used in fluid analysis [38].

Figure 27.4 shows a nano-particle induced in plasmonic oscillation by a light wave. Figure 27.5 shows that different color fluids can be obtained by immersing nano-particles in fluids with different background permittivity (ε in (27.3.6) and (27.3.7)) causing the plasmonic particles to resonate at different frequencies. This is because the resonance frequency of the plasmonic nanoparticle is obtained by solving $\varepsilon_s(\omega) + 2\varepsilon = 0$, which depends on the background medium, ε.

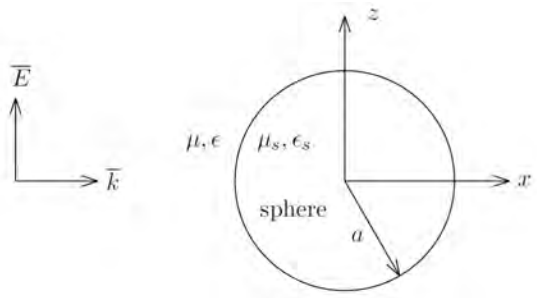


Figure 27.3: A plane electromagnetic wave incident on a particle. When the particle size is small compared to wavelength, electrostatic analysis can be used to solve this problem (courtesy of Kong [34]).

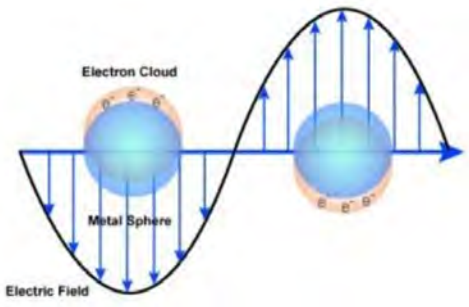


Figure 27.4: A nano-particle undergoes electromagnetic oscillation when an electromagnetic wave impinges on it. The oscillation is inordinately large when the incident wave's frequency coincides with the resonance frequency of the plasmonic particle (picture courtesy of sigmaaldric.com).

Figure 27.5: Different color fluids containing nano-particles can be obtained by changing the permittivity ε of the background fluids in (27.3.6) and (27.3.7) (courtesy of nanocomposix.com).

27.3.3 Quasi-Static Electromagnetic Theory–Amplification of the Flux Terms

In closing, we would like to make one more remark. The right-hand side of (27.3.8), is the integral form of Faraday's law, is essential for capturing the physical mechanism of an inductor and flux linkage. And yet, if we drop it, there will be no inductors in this world. To examine it further,

$$\oint_C \mathbf{E}' \cdot d\mathbf{l} = -j\omega\mu_0 \frac{L}{\eta_0} \iint_S d\mathbf{S} \cdot \mathbf{H} \tag{27.3.8}$$

In the inductor, the right-hand side has been amplified by multiple turns, effectively increasing S, the flux linkage area. Or one can think of an inductor as having a much longer effective length L_{eff} when untwined so as to compensate for decreasing frequency ω. Hence, the importance of flux linkage or the inductor in Faraday's law is not diminished unless $\omega = 0$.

By the same token, displacement current in Ampere's law can be enlarged by using capacitors. In this case, even when no electric current $\mathbf{J}$ flows through the capacitor, displacement current flows and the generalized Ampere's law becomes

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = j\omega\varepsilon\eta_0 L \iint_S d\mathbf{S} \cdot \mathbf{E}' \tag{27.3.9}$$

The right-hand side can be enlarged by making S large to amplify the displacement current. Thus, the displacement current in a capacitor cannot be ignored unless $\omega = 0$. Therefore, when $\omega \neq 0$, or in quasi-static case, inductors and capacitors in circuit theory are extremely important, because they amplify the flux linkage and the displacement current effects, as we shall study next. In summary, the full physics of Maxwell's equations is not lost in circuit theory: the induction term in Faraday's law, and the displacement current in Ampere's law are still retained.[7] We can still have wave phenomena in circuit theory as exemplified by the lumped element transmission-line model. In fact, by enlarging the line capacitance and line inductance, the phase velocity of the

[7]Putatively, Maxwell got his epiphany to add displacement current to Ampere's law when he studied current through a capacitor.

wave on such a line can be reduced making it into a slow-wave structure. That explains the success of circuit theory in electromagnetic engineering! Circuit theory bears the burden of electromagnetic engineering down to the chip level. As mentioned before, the full glory of Maxwell's equations is not lost in circuit theory by the use of capacitors and inductors.

27.4 Helmholtz Decomposition

Given the proof of uniqueness of a field when its curl and its divergence are specified (see Section 4.1.1), we shall review Helmholtz decomposition next. Helmholtz decomposition states that [187, 101, 49] an arbitrary field $\mathbf{A}$ can be written as a sum of curl-free field and a divergence-free field. In other words,

$$\mathbf{A} = \mathbf{A}_{\parallel} + \mathbf{A}_{\perp} = -\nabla\phi + \nabla \times \mathbf{P} \tag{27.4.1}$$

where $\mathbf{A}_{\parallel} = -\nabla\phi$ is curl-free and $\mathbf{A}_{\perp} = \nabla \times \mathbf{P}$ is divergence-free. Curl-free is variously known as irrotational (or longitudinal) and divergence-free is referred to as solenoidal (or transverse).[8]

Since a vector can be uniquely defined by its divergence $\nabla \cdot \mathbf{A} = s$ and its curl $\nabla \times \mathbf{A} = \mathbf{c}$, taking the divergence and curl of (27.4.1), we have

$$\nabla \cdot \mathbf{A} = -\nabla^2\phi = s \tag{27.4.2}$$

$$\nabla \times \mathbf{A} = \nabla \times \nabla \times \mathbf{P} = \mathbf{c} \tag{27.4.3}$$

where s is the source for Poisson equation, and $\mathbf{c}$ is the source for $\mathbf{P}$. We proceed with the proof of Helmholtz theorem by deriving the expression for ϕ and $\mathbf{P}$ and relate them to the sources s and $\mathbf{c}$.

The solution of $\mathbf{P}$ in (27.4.3) is not unique, unless we set $\nabla \cdot \mathbf{P} = 0$, and (27.4.3) becomes vector Poisson equation with a unique solution (see (4.1.10)). The solution to (27.4.2) and (27.4.3) are then well known using static Green's function method as in (4.1.11):

$$\phi(\mathbf{r}) = \frac{1}{4\pi} \int \frac{s(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \tag{27.4.4}$$

$$\mathbf{P}(\mathbf{r}) = \frac{1}{4\pi} \int \frac{\mathbf{c}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}'. \tag{27.4.5}$$

27.5 Mode Decomposition in a General Cavity

Helmholtz decomposition can also be derived from the decomposition of cavity fields in terms of cavity modes. As shall be seen, for a general cavity, the field inside can be decomposed into the sum of divergence-free field and curl-free field. We can couch the problem into one in finding the eigenvector (eigenmodes) that will solve the following eigenvalue problem

$$-\nabla^2\mathbf{A} = \omega^2\mu\epsilon\mathbf{A} \rightarrow -\nabla^2\mathbf{A} = \lambda\mathbf{A} \tag{27.5.1}$$

[8]It should be noted that the term longitudinal and transverse are first used in the context of plane wave in Fourier space, where $\nabla \times \mathbf{A} = 0 \rightarrow \mathbf{k} \times \mathbf{A} = 0$ and $\nabla \cdot \mathbf{A} = 0 \rightarrow \mathbf{k} \cdot \mathbf{A} = 0$, (where k is the wave vector or propagation direction of the plane wave), suggesting that the longitudinal (transverse) field is parallel (perpendicular) to the propagation direction in the Fourier space. Here, we use these terms for curl-free and divergence-free in general cases, with a caveat that they do not have the same physical meaning as in plane wave.

With appropriate boundacy conditions, it can be shown that the Laplacian operator is a negative definite, and a Hermitian operator. So the left-hand side is always positive definite with real eigenvalues. When the above equation is converted into a matrix eigenvalue problem, $\overline{\overline{\mathbf{W}}} \cdot \mathbf{A} = \lambda \mathbf{A}$, with $\overline{\overline{\mathbf{W}}}$ being Hermitian and positive definite. The eigenvalues will be always positive and real.[9] Because of this, the eigenvectors are complete, and the eigenvalues λ are real.

Now we can rewrite the above EVP as

$$\nabla \times \nabla \times \mathbf{A} - \nabla \nabla \cdot \mathbf{A} = \lambda \mathbf{A}. \tag{27.5.2}$$

It is the same EVP but with the left-hand side written differently. The above is just (27.2.24) with the right-hand side set to zero, or we are looking for the homogeneous solution of (27.2.24), which are the eigenmodes or the modal solutions.

First, we decompose the above into two EVP's, namely,

$$\nabla \times \nabla \times \mathbf{A}_a = \lambda_a \mathbf{A}_a, \text{with eigenvalue } \lambda_a \tag{27.5.3}$$

$$-\nabla \nabla \cdot \mathbf{A}_b = \lambda_b \mathbf{A}_b, \text{with eigenvalue } \lambda_b \tag{27.5.4}$$

From the above equations, we can show easily that $\nabla \cdot \mathbf{A}_a = 0$ and $\nabla \times \mathbf{A}_b = 0$. It can be shown that these modes of (27.5.3) and (27.5.4) are also the modes of (27.5.2). Furthermore, these modes are ortogonal and form a complete set [188]. Therefore, an arbitrary field of a cavity can be decomposed into sum of modes from (27.5.2) as well as sum of div-free modes and curl-free modes from (27.5.3) and (27.5.4). This is also the statement of Helmholtz decomposition.

27.5.1 Excitation of Cavity Field with Eigenmode Expansion

When a source is in a cavity, the field in the cavity will be excited by the source. This can be demonstrated with setting the right-hand side of (27.2.10) to be nonzero, and finding the field in a cavity. And hence, we need to find the inhomogeneous solution of the problem.[10] It is best that we illustrate the solution to this problem using matrix-algebra notation.[11]

Consider the equation (27.2.10) which is

$$(\nabla^2 + \beta^2)\mathbf{A}(\mathbf{r}) = -\mu \mathbf{J}(\mathbf{r}) \tag{27.5.5}$$

The above partial differential equation is analogous to the matrix equation

$$\overline{\mathbf{L}} \cdot \mathbf{f} = \mathbf{s} \tag{27.5.6}$$

The matrix operator $\overline{\mathbf{L}}$ is analogous to $(\nabla^2 + \beta^2)$ and the vector $\mathbf{f}$ is analogous to the vector function $\mathbf{A}(\mathbf{r})$, and $\mathbf{s}$ is analogous to $-\mu \mathbf{J}(\mathbf{r})$. Assume that we have already found the eigenmodes $\mathbf{v}_i$ of a very similar EVP (such as in (27.5.1)),

$$\overline{\mathbf{L}}_0 \cdot \mathbf{v}_i = \beta_i^2 \mathbf{v}_i \tag{27.5.7}$$

[9] We shall learn how to convert this equation into a matrix equation later in the course.
[10] This is also the solution with source excitation which is a very common electromagnetics problem.
[11] These notations turn out to be most complete and comprehensive.

where the matrix operator $\overline{\mathbf{L}}_0$ is analogous to $-\nabla^2$ operator in (27.5.1), and similarly, $\mathbf{v}_i$ is analogous to the vector field $\mathbf{A}(\mathbf{r})$, β_i^2 is analogous to to the eigenvalue λ. So the above equation is analogous to the equation

$$-\nabla^2 \mathbf{A}_i(\mathbf{r}) = \beta_i^2 \mathbf{A}_i(\mathbf{r}) \tag{27.5.8}$$

where the subscript i is to emphasize the fact that this is an eigenvalue problem with eigenfunction $\mathbf{A}_i(\mathbf{r})$ and eigenvalue β_i^2. Here, the eigenfunction is analogous to the eigenvector.

Assume that the problem we are required to solve is of the form

$$\overline{\mathbf{L}} \cdot \mathbf{f} = \left(\overline{\mathbf{L}}_0 + \beta_0^2 \overline{\mathbf{I}}\right) \cdot \mathbf{f} = \mathbf{s} \tag{27.5.9}$$

where $\overline{\mathbf{L}} = \overline{\mathbf{L}}_0 + \beta_0^2 \overline{\mathbf{I}}$. The above is analogous to (27.2.24) with the source term $\mathbf{s}$ equivalent to $-\mu \mathbf{J}(\mathbf{r})$. Then we can expand the field $\mathbf{f}$ in terms of the eigenmodes (or eigenfunctions) we have found in (27.5.7). These eigenmodes or eigenfunctions form an orthonormal and complete set, since the operator $-\nabla^2$ is analogous to a Hermitian matrix operator which has a complete set of eigenvectors.

Namely, now we let

$$\mathbf{f} = \sum_i a_i \mathbf{v}_i \tag{27.5.10}$$

Then

$$\overline{\mathbf{L}} \cdot \mathbf{f} = \overline{\mathbf{L}} \cdot \sum_i a_i \mathbf{v}_i = \sum_i a_i \left(\overline{\mathbf{L}}_0 + \beta_0^2 \overline{\mathbf{I}}\right) \cdot \mathbf{v}_i = \sum_i a_i \left(\beta_i^2 + \beta_0^2\right) \cdot \mathbf{v}_i = \mathbf{s} \tag{27.5.11}$$

In order to solve for the unknowns a_i, we use mode orthogonality, which can be orthonormalized. In a word, we have

$$\mathbf{v}_j^\dagger \cdot \mathbf{v}_i = \delta_{j,i} \tag{27.5.12}$$

Then after inner-product the (27.5.11) above with $\mathbf{v}_j^\dagger$ from the left, we have

$$\mathbf{v}_j^\dagger \cdot \overline{\mathbf{L}} \cdot \mathbf{f} = \sum_i a_i (\beta_i^2 + \beta_0^2) \mathbf{v}_j^\dagger \cdot \mathbf{v}_i = \mathbf{v}_j^\dagger \cdot \mathbf{s} \tag{27.5.13}$$

Using mode orthonormality, we can solve the above to obtain

$$a_j = \frac{\mathbf{v}_j^\dagger \cdot \mathbf{s}}{\beta_j^2 + \beta_0^2} \tag{27.5.14}$$

When a_j is used in (27.5.10), the above is the general eigenmode expansion solution of a cavity excited by a source. When the source is converted to a point source, it is the Green's function of the system. Since $\mathbf{v}_i$ consists of div-free and curl-free modes, the Helmholtz decomposition of the field inside a cavity excited by a general source is also obvious.

Exercises for Lecture 27

Problem 27-1: The equations for vector and scalar potentials have been found using the Lorenz gauge in the lecture notes.

(i) Now, find these equations using Coulomb's gauge. **Hint: This is described in J.D. Jackson's book, and many other physics texts.**

(ii) Verify by back substitution that (27.2.28) is in fact a solution of (27.2.27).

(iii) Describe how you would use the electric vector potential and the magnetic scalar potential to find the electromagnetic field if you only have magnetic current and magnetic charge as sources.

(iv) Explain when the static electromagnetic theory can be used to approximate a time-varying electromagnetic field.

(v) Derive (27.5.14), and explain why Helmholtz decomposition is obvious in (27.5.10) once the cavity excitation problem has been solved.

Chapter 28

Radiation by a Hertzian Dipole

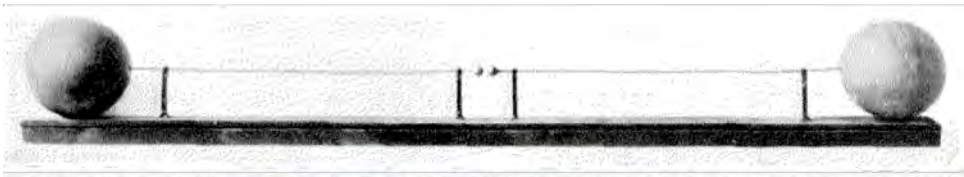
Radiation of electromagnetic field is of ultimate importance for wireless communication systems. The first demonstration of the wave nature of electromagnetic field was by Heinrich Hertz in 1888 [19], some 23 years after Maxwell's equations were fully established. Guglielmo Marconi, after much dogged perserverence with a series of experiments, successfully transmitted wireless radio signal from Cornwall, England to Newfoundland, Canada in 1901 [189]. The experiment was serendipitous since he did not know that the ionosphere was on his side: The ionosphere helped to bounce the radio wave back to earth from outer space. Marconi's success ushered in the age of wireless communication, which is omni-present in our present daily lives. Hence, radiation by arbitrary sources is an important topic for antennas and wireless communications. We will start with studying the Hertzian dipole which is the simplest of radiation sources we can think of.

28.1 History

The original historic Hertzian dipole experiment is shown in Figure 28.1. It was done in 1887 by Heinrich Hertz [19]. The schematic for the original experiment is also shown in Figure 28.2.

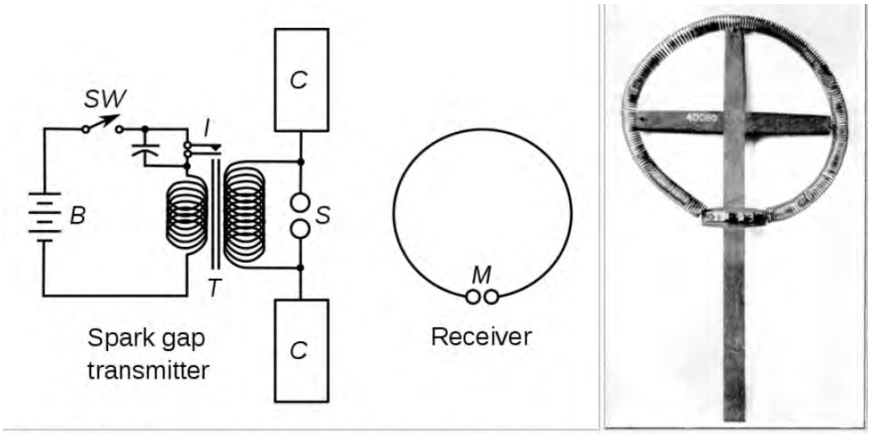
A metallic sphere has a capacitance which can be found in closed form with respect to infinity or a ground plane.[1] Hertz could use those knowledge to estimate the mutual capacitance of the two spheres, and also, he could estimate the inductance of the leads that were attached to the dipole, therefore, the resonance frequency of his antenna can be calculated. The large sphere is needed to have a large capacitance, so that current can be driven through the wires. As we shall see, the radiation strength of the dipole is proportional to $p = ql$, the dipole moment. Here, q is the charge at the ends of the dipole, and l is its length.

[1] We shall learn later that this problem can be solved in closed form using image theorem.



Hertz's first radio transmitter: a dipole resonator consisting of a pair of one meter copper wires with a 7.5 mm spark gap between them, ending in 30 cm zinc spheres.[12] When an induction coil applied a high voltage between the two sides, sparks across the spark gap created standing waves of radio frequency current in the wires, which radiated radio waves. The frequency of the waves was roughly 50 MHz, about that used in modern television transmitters.

Figure 28.1: Hertz's original experiment on a small dipole (courtesy of Wikipedia [19]).



Hertz's 1887 apparatus for generating and detecting radio waves: a spark transmitter *(left)* consisting of a dipole antenna with a spark gap *(S)* powered by high voltage pulses from a Ruhmkorff coil *(T)*, and a receiver *(right)* consisting of a loop antenna and spark gap.

One of Hertz's radio wave receivers: a loop antenna with an adjustable micrometer spark gap *(bottom)*.[12]

Figure 28.2: More on Hertz's original experiment on a small dipole. The antenna was powered by a transformer. The radiated electromagnetic field was detected by a loop receiver antenna that generated a spark at its gap M (courtesy of Wikipedia [19]).

28.2 Approximation by a Point Source

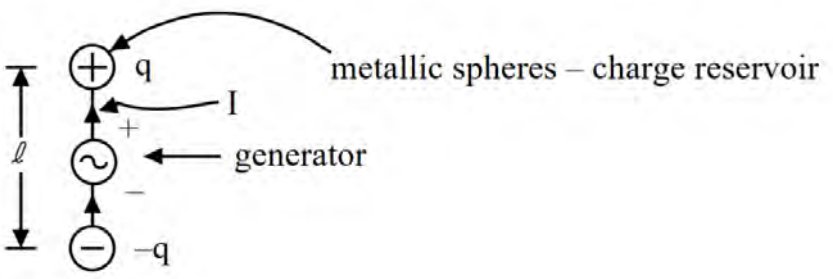


Figure 28.3: Schematic of a small Hertzian dipole which is a close approximation of that first proposed by Hertz.

Figure 28.3 is the schematic of a small Hertzian dipole resembling the original dipole that Hertz made. Assuming that the spheres at the ends store charges of value q, and l is the effective length of the dipole, then the dipole moment $p = ql$. The charge q is varying time-harmonically because it is driven by the generator. Since

$$\frac{dq}{dt} = I,$$

we have the current moment

$$Il = \frac{dq}{dt} l = j\omega ql = j\omega p \tag{28.2.1}$$

for this Hertzian dipole.

A Hertzian dipole is a dipole which is much smaller than the wavelength under consideration so that we can approximate it by a point current distribution, or a current density. Mathematically, it is given by [47, 34]

$$\mathbf{J}(\mathbf{r}) = \hat{z} Il\delta(x)\delta(y)\delta(z) = \hat{z} Il\delta(\mathbf{r}) \tag{28.2.2}$$

The dipole is as shown in Figure 28.3 schematically. As long as we are not too close to the dipole so that it does not look like a point source anymore, the above is a good mathematical model and approximation for describing a Hertzian dipole.

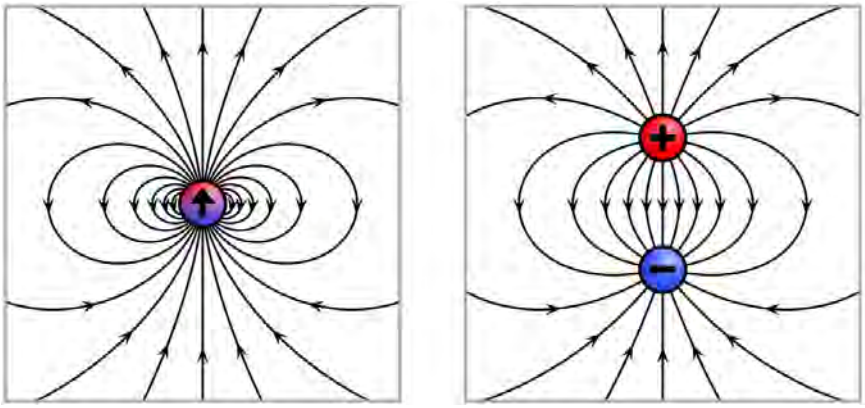


Figure 28.4: The field of a point dipole field where the separation of the two charges is infinitesimally small versus that of a dipole field with a finite separation between the charges. When one is far away from the dipole sources, their fields are similar to each other (courtesy of Wikepedia).

We have learnt previously that the vector potential is related to the current as follows:

$$\mathbf{A}(\mathbf{r}) = \mu \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{28.2.3}$$

Since the current is a 3D delta function in space, using the sifting property of a delta function, and that $\delta(\mathbf{r}) = \delta(x)\delta(y)\delta(y)$, the above integral can be evaluated in closed form. The corresponding vector potential is then given by

$$\mathbf{A}(\mathbf{r}) = \hat{z} A_z = \hat{z} \frac{\mu I l}{4\pi r} e^{-j\beta r} \tag{28.2.4}$$

Since the vector potential $\mathbf{A}(\mathbf{r})$ is cylindrically symmetric (also called axi-symmetric or axially symmetric), the corresponding magnetic field is obtained by using that $\mathbf{B} = \nabla \times \mathbf{A}$. Thus, using the curl operator in cylindrical coordinates,

$$\mathbf{H} = \frac{1}{\mu} \nabla \times \mathbf{A} = \frac{1}{\mu} \left(\hat{\rho} \frac{1}{\rho} \frac{\partial}{\partial \phi} A_z - \hat{\phi} \frac{\partial}{\partial \rho} A_z \right) \tag{28.2.5}$$

Due to axi-symmetry, then $\frac{\partial}{\partial \phi} = 0$. Here, $r = \sqrt{\rho^2 + z^2}$. In the above, we can use the chain rule that

$$\frac{\partial}{\partial \rho} = \frac{\partial r}{\partial \rho} \frac{\partial}{\partial r} = \frac{\rho}{\sqrt{\rho^2 + z^2}} \frac{\partial}{\partial r} = \frac{\rho}{r} \frac{\partial}{\partial r}$$

to find the $\hat{\phi}$ component of $\mathbf{H}$, since A_z is a function of r only. As a result,

$$\mathbf{H} = -\hat{\phi} \frac{\rho}{r} \frac{Il}{4\pi} \left(-\frac{1}{r^2} - j\beta \frac{1}{r} \right) e^{-j\beta r} \tag{28.2.6}$$

In spherical coordinates, $\frac{\rho}{r} = \sin\theta$, and (28.2.6) becomes [34]

$$\mathbf{H} = \hat{\phi} H_\phi = \hat{\phi} \frac{Il}{4\pi r^2} (1 + j\beta r) e^{-j\beta r} \sin\theta \tag{28.2.7}$$

Notice that H_ϕ is a function of r and θ only in spherical coordinates. The electric field can be derived using Maxwell's equations in spherical coordinates, viz.,

$$\begin{aligned} \mathbf{E} &= \frac{1}{j\omega\epsilon} \nabla \times \mathbf{H} = \frac{1}{j\omega\epsilon} \left(\hat{r} \frac{1}{r \sin\theta} \frac{\partial}{\partial\theta} \sin\theta H_\phi - \hat{\theta} \frac{1}{r} \frac{\partial}{\partial r} r H_\phi \right) \\ &= \frac{Ile^{-j\beta r}}{j\omega\epsilon 4\pi r^3} \left[\hat{r} 2\cos\theta (1 + j\beta r) + \hat{\theta} \sin\theta (1 + j\beta r - \beta^2 r^2) \right] \end{aligned} \tag{28.2.8}$$

The above expression is rather complicated and it is hard to elucidate the physics from the math (or we cannot see the trees in the forest). However, we can see that as $r \to \infty$, there are terms that decay as $1/r^3$, $1/r^2$, and $1/r$. The complex Poynting's vector is proportional to $\mathbf{E} \times \mathbf{H}^*$, and the term that will convect power to infinity has to decay as $1/r^2$ for energy conservation, since the energy is radiating into a spherical surface whose area grows as r^2.

Terms that decay faster than $1/r^2$ cannot convect (carry) energy to infinity, but store reactive power that can exchange with the power from the source. As in the case of a capacitor or an inductor, the reactive power takes energy from the source, and then return it to the source again.

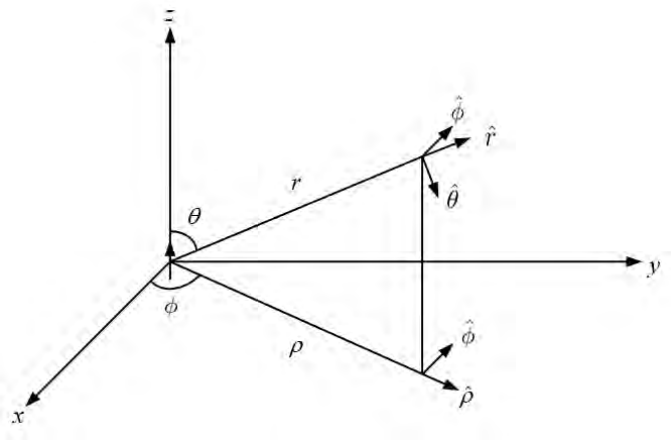


Figure 28.5: Spherical coordinates are used to calculate the fields of a Hertzian dipole.

28.2.1 Case I. Near Field, $\beta\mathbf{r} \ll \mathbf{1}$

As we have seen, the yardstick in electromagnetics is the wavelength. Since $\beta = \frac{\omega}{c} = \frac{2\pi}{\lambda}$, $\beta r \ll 1$ implies that one is close to the dipole source in terms of wavelength. We are in the near field of the antenna. In the near field of an antenna, we have mainly reactive power: it is characterized by that the energy flows from the source to the medium in one cycle, and the converse in another cycle.

Since $\beta r \ll 1$,[2] retardation effect within this short distance from the point dipole can be ignored. Also, we let $\beta r \to 0$, and keeping the largest terms (or leading order terms in math parlance), then from (28.2.8), with $Il = j\omega p$, we have

$$\mathbf{E} \cong \frac{p}{4\pi\epsilon r^3}(\hat{r}2\cos\theta + \hat{\theta}\sin\theta), \qquad \beta r \ll 1 \tag{28.2.9}$$

For the $\mathbf{H}$ field, from (28.2.7), with $\beta r \ll 1$, then

$$\mathbf{H} \cong \hat{\phi}\frac{j\omega p}{4\pi r^2}\sin\theta \tag{28.2.10}$$

or

$$\eta_0\mathbf{H} \cong \hat{\phi}\frac{j\beta r p}{4\pi\varepsilon r^3}\sin\theta \tag{28.2.11}$$

Now comparing (28.2.11) and (28.2.9), we see that (28.2.11) is $O(\beta r)$ smaller than (28.2.9). Since $\mathbf{E}$ and $\eta_0\mathbf{H}$ have the same unit, we can now compare them. Thus, it is seen that

$$\eta_0\mathbf{H} \ll \mathbf{E}, \qquad \text{when } \beta r \ll 1 \tag{28.2.12}$$

where $p = ql$ is the dipole moment.[3] The above implies that in the near field, the electric field dominates over the magnetic field. Moreover, the $\mathbf{E}$ and the $\mathbf{H}$ fields are out of phase, meaning that the power that corresponds to a complex Poynting's vector $\mathbf{E}\times\mathbf{H}^*$ is almost purely imaginary or reactive (see Section 10.3.1). This reactive power pumps stored energy into the near field, and siphons energy from the near field as well (similar to a capacitor).

In the above, βr could be made very small by making $\frac{r}{\lambda}$ small or by making $\omega \to 0$. The above becomes like the static field of a dipole. Another viewpoint is that in the near field, the field varies rapidly, and space derivatives are much larger than the time derivative.[4]

For instance,

$$\frac{\partial}{\partial x} \gg \frac{\partial}{c\partial t}$$

Alternatively, we can say that the above is equivalent to

$$\frac{\partial}{\partial x} \gg \frac{\omega}{c}, \qquad \frac{\partial^2}{\partial x^2} \gg \frac{\omega^2}{c^2}, \qquad \frac{\partial^2}{\partial x^2} \gg \frac{1}{c^2}\frac{\partial^2}{\partial t^2}$$

or that

$$\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} \approx \nabla^2$$

which is just the Laplacian operator. In other words, static theory prevails over dynamic theory when $\beta r \ll 1$. The above approximations are consistent with that the retardation effect is negligible over this length scale.

[2]Please be reminded that β and k are synonymous in this course: β for microwave engineers, and k for optical engineers.

[3]Here, $\eta_0 = \sqrt{\mu/\varepsilon}$. We multiply $\mathbf{H}$ by η_0 so that the quantities we are comparing have the same unit.

[4]This is in agreement with our observation that electromagnetic fields are great contortionists: They will deform themselves to match the boundary first before satisfying Maxwell's equations. Since the source point is very small, the fields will deform themselves so as to satisfy the boundary conditions near to the source region. If this region is small compared to wavelength, the fields will vary rapidly and spatially over a small length scale compared to wavelength.

28.2.2 Case II. Far Field (Radiation Field), $\beta \mathbf{r} \gg 1$

This is an interesting zone where we can see electromagnetics power being convected (carried) to infinity from the source. This is also known as the far zone.[5] In this case, electromagnetics retardation effect is important. In other words, phase delay cannot be ignored. Thus keeping the leading order terms in (28.2.8), we then have

$$\mathbf{E} \cong \hat{\theta} j\omega\mu \frac{Il}{4\pi r} e^{-j\beta r} \sin\theta \tag{28.2.13}$$

and similarly from (28.2.7)

$$\mathbf{H} \cong \hat{\phi} j\beta \frac{Il}{4\pi r} e^{-j\beta r} \sin\theta \tag{28.2.14}$$

Note that $\frac{E_\theta}{H_\phi} = \frac{\omega\mu}{\beta} = \sqrt{\frac{\mu}{\epsilon}} = \eta_0$ which is similar to the intrinsic impedance relationship of a plane wave. Here, $\mathbf{E}$ and $\mathbf{H}$ are orthogonal to each other and they are both orthogonal to the direction of propagation, as in the case of a plane wave. Or in a word, a spherical wave resembles a plane wave in the far field approximation.

The *radiation field pattern* of a Hertzian dipole is the plot of $|\mathbf{E}|$ as a function of θ at a constant $\mathbf{r}$ when $\beta r \gg 1$ or in the far field of the antenna. Hence, it is proportional to $\sin\theta$, and it can be proved that it is a circle as shown in Figure 28.6. Also, notice that in the above, the $\mathbf{E}$ and the $\mathbf{H}$ are in phase in the far field, meaning that the complex Poynting's vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}^*$ is purely real, and that the $\mathbf{S}$ decays as $1/r^2$. This implies energy conservation as the wave propagates, and the power can travel to infinity.

28.3 Radiation Power of a Hertzian Dipole

The time average power flow in the far field, after using (28.2.13) and (28.2.14), is given by

$$\langle \mathbf{S} \rangle = \frac{1}{2} \Re e[\mathbf{E} \times \mathbf{H}^*] = \hat{r} \frac{1}{2} \eta_0 |H_\phi|^2 = \hat{r} \frac{\eta_0}{2} \left(\frac{\beta Il}{4\pi r} \right)^2 \sin^2\theta = \hat{r} \langle S_r \rangle \tag{28.3.1}$$

$$\langle S_r \rangle = \frac{\eta_0}{2} \left(\frac{\beta Il}{4\pi r} \right)^2 \sin^2\theta \tag{28.3.2}$$

The *radiation power pattern* is the plot of $r^2 \langle S_r \rangle$ as $r \to \infty$ or in the far field as shown in Figure 28.7.

[5] This is also called the Fraunhofer zone in German.

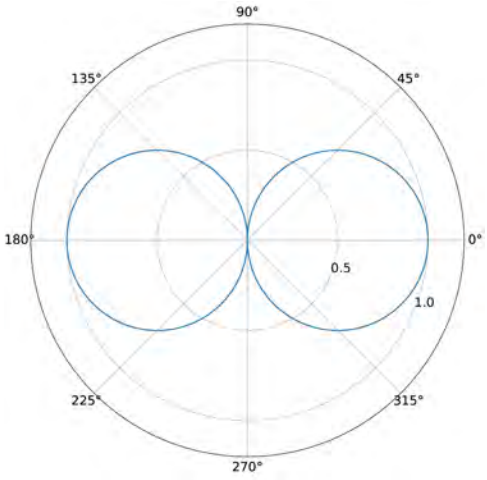


Figure 28.6: Radiation field pattern of a Hertzian dipole. It can be shown that the pattern is a circle.

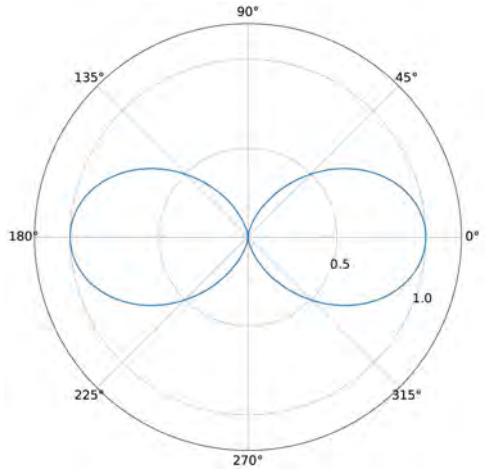


Figure 28.7: Radiation power pattern of a Hertzian dipole which is also the same as the directive gain pattern.

The total power radiated by a Hertzian dipole is thus given by

$$P_T = \int d\Omega \langle S_r \rangle = \int_0^{2\pi} d\phi \int_0^{\pi} d\theta r^2 \sin\theta \langle S_r \rangle = 2\pi \int_0^{\pi} d\theta \frac{\eta_0}{2} \left(\frac{\beta I l}{4\pi} \right)^2 \sin^3\theta \tag{28.3.3}$$

Since

$$\int_0^{\pi} d\theta \sin^3\theta = -\int_1^{-1} (d\cos\theta)[1-\cos^2\theta] = \int_{-1}^{1} dx(1-x^2) = \frac{4}{3} \tag{28.3.4}$$

then the total power becomes

$$P_T = \frac{4}{3}\pi\eta_0 \left(\frac{\beta I l}{4\pi}\right)^2 = \frac{\eta_0(\beta I l)^2}{12\pi} \tag{28.3.5}$$

28.3.1 Radiation Resistance–Circuit Equivalence of a Hertzian Dipole

Engineers love to replace complex systems with simpler systems. Simplicity rules again! For example, the voltage or current sources driving an antenna is usually made from electronic circuits. Hence, it will be expedient to replace an antenna with circuit equivalence so that it can interface with the driving circuit components. A raw Hertzian dipole, when driven by a voltage source, essentially looks like a capacitor due to the preponderance of electric field energy stored in the dipole field. But at the same time, the dipole radiates giving rise to radiation loss. Thus a simple circuit equivalence of a Hertzian dipole is a capacitor in series with a resistor. The resistor accounts for radiation loss of the dipole.

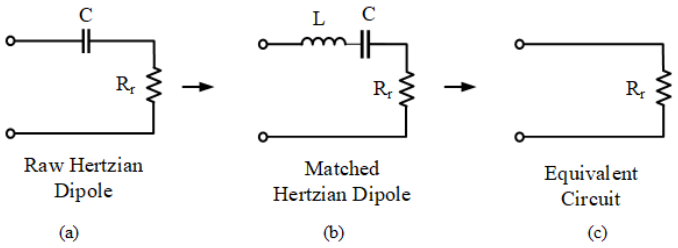


Figure 28.8: A short Hertzian dipole, shown in Figure 28.1, resembles an open circuit with a capacitor connected to it because the two spheres shown in Figure 28.1 will have charges stored in them giving rise to electric energy stored. The radiation loss of the antenna can be modeled by a radiation resistance R_r. Hence, we have (a) an equivalent circuit of a raw Hertzian dipole without matching; (b) an equivalent circuit of a matched Hertzian dipole (using maximum power transfer theorem [55, Sections 2.5 and 4.3], [190]); (c) an equivalent circuit of a matched dipole at the resonance frequency of the LC tank circuit.

Hence, the way to drive the Hertzian dipole effectively is to use matching network according to maximum power transfer theorem.[6] Or an inductor has to be added in series with the intrinsic capacitance of the Hertzian dipole to cancel it at the resonance frequency of the tank circuit.

[6]The maximum power transfer theorem has to be used with caution, because with the recent advances in power electronics and transformers, the limit of maximum power transfer theorem can be exceeded.

Eventually, after matching, the Hertzian dipole can be modeled as just a resistor. Then the power absorbed by the Hertzian dipole from the driving source is $P_T = \frac{1}{2}I^2R_r$. Thus, the *radiation resistance* R_r is the effective resistance that will dissipate the same power as the total radiation power P_T when a current I flows through the resistor. It is defined by [34]

$$R_r = \frac{2P_T}{I^2} = \eta_0 \frac{(\beta l)^2}{6\pi} \approx 20(\beta l)^2, \qquad \text{where } \eta_0 = 377 \approx 120\pi\ \Omega \tag{28.3.6}$$

In the above, we have used (28.3.5) for the total power radiated by the Hertzian dipole. For example, for a Hertzian dipole with $l = 0.1\lambda$, $R_r \approx 8\Omega$.

The above assumes that the current is uniformly distributed over the length of the Hertzian dipole. This is true if there are two charge reservoirs at its two ends, as in Hertz original dipole antenna. For a small dipole with no charge reservoir at the two ends, the currents have to vanish at the tips of the dipole as shown in Figure 28.9. The effective length of an equivalent Hertzian dipole for the dipole with triangular distribution is *half* of its actual length due to the manner the currents are distributed.[7] Such a formula can be used to estimate the radiation resistance of a small/short dipole.

For a filamental wire current with a non-uniform current, it can be approximated with a current given by

$$\mathbf{J}(\mathbf{r}) = \hat{z}\delta(x)\delta(y)I(z) \tag{28.3.7}$$

where $I(z)$ is the current distribution on the filamental wire. When such a current is used in the integral in (28.2.3), the integral does not have a closed form solution, unless we make further approximations like ignoring the phase term $\exp(-j\beta|\mathbf{r}-\mathbf{r}'|$ which accounts for retardation effect. Then (28.2.3), if $\mathbf{r}'$ is assume to be small always, can be approximated as

$$\mathbf{A}(\mathbf{r}) \cong \mu \iiint_V d\mathbf{r}'\mathbf{J}(\mathbf{r}')\frac{e^{-j\beta|\mathbf{r}|}}{4\pi|\mathbf{r}|} = \mu\frac{e^{-j\beta r}}{4\pi r}\iiint_V d\mathbf{r}'\mathbf{J}(\mathbf{r}') \tag{28.3.8}$$

where the phase term has been taken outside the integral. Then the radiation field is proportional to the integral summation of the radiation current alone.

For example, a half-wave dipole does not have a triangular current distribution but a sinusoidal one as shown in Figure 28.10. Nevertheless, we approximate the current distribution of a half-wave dipole with a triangular distribution, and apply the above formula (28.3.6). Picking $a = \frac{\lambda}{2}$, and letting $l_{\text{eff}} = \frac{\lambda}{4}$ in (28.3.6), we have

$$R_r \approx 50\Omega \tag{28.3.9}$$

[7] As shall be shown, when the dipole is short, the details of the current distribution is inessential in determining the radiation field. It is the area under the current distribution that is important.

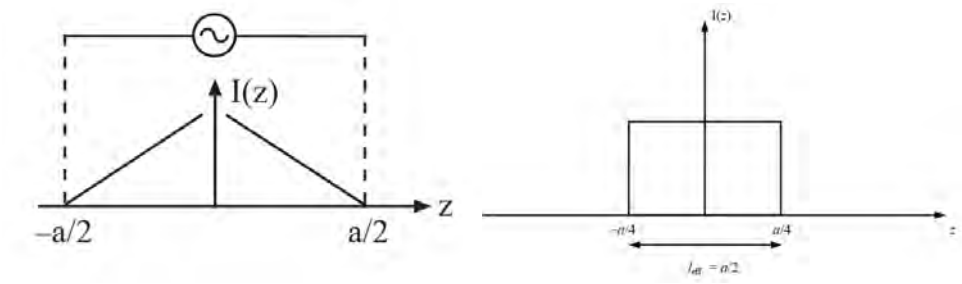


Figure 28.9: The current pattern on a short dipole can be approximated by a triangle since the current has to vanish at the end points of the short dipole. Furthermore, this dipole can be approximated by an effective Hertzian dipole half its length with uniform current. A more accurate current distribution on a wire dipole antenna can be found by using numerical methods, such as the method of moments [191, 192, 193]

The true current distribution on a half-wave dipole resembles that shown in Figure 28.10. The current is zero at the end points, but the current has a more sinusoidal-like distribution as in a transmission line. Hence, a half-wave dipole is not much smaller than a wavelength and does not qualify to be a Hertzian dipole. Furthermore, the current distribution on the half-wave dipole is not triangular in shape as above. Moreover, when we calculate (28.2.4), we assume that there is no phase delay between different parts of the current.[8] This is not true when the dipole antenna is not short compared to wavelength. This retardation effect, that comes from the $\exp(-j\beta|\mathbf{r}-\mathbf{r}'|)$ in (28.2.3) has to be accounted for. The calculation of such radiation integrals will be discussed in the next lecture.

A more precise calculation using semi-analytic method shows that $R_r = 73\ \Omega$ for a half-wave dipole [57][p. 332].[9] This also implies that a half-wave dipole with sinusoidal current distribution is a better radiator than a dipole with just a triangular current distribution. More precise value can be obtained using numerical methods.

In fact, one can think of a half-wave dipole as a flared, open transmission line. In the beginning, this flared open transmission line came in the form of biconical antennas which are shown in Figure 28.11 [194]. If we recall that the characteristic impedance of a transmission line is $\sqrt{L/C}$, then as the spacing of the two metal pieces becomes bigger, the equivalent characteristic impedance becomes bigger. Therefore, the impedance can gradually transform from a small impedance like 50 Ω to that of free space, which is 377 Ω. This impedance matching helps mitigate reflection from the ends of the flared transmission line, and enhances radiation. Because of the matching nature of bicone antennas, they are better radiators with higher radiaton loss and lower Q. Thus they have a broader bandwidth, and are important in UWB (ultra-wide band) antennas [195].

[8] A more precise calculation of the current distribution requires a numerical method like the method of moments.
[9] This explains why the characteristic impedance of some transmission line is 75 Ω.

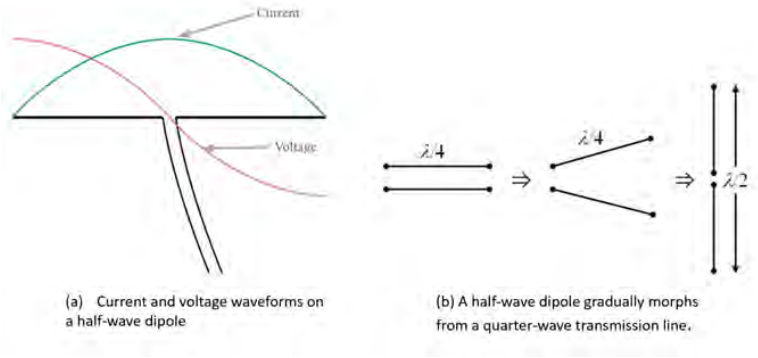


Figure 28.10: Approximate current distribution on a half-wave dipole (courtesy of electronics-notes.co). The currents are zero at the two end tips due to the current continuity equation, or KCL. A more precise calculation of its input impedance is given in [57].

Figure 28.11: A bicone antenna can be thought of as a flared transmission line with gradually changing characteristic impedance. This enhances impedance matching and the radiation of the antenna (courtesy of antennasproduct.com).

Exercises for Lecture 28

Problem 28-1:

(i) Derive (28.2.7) and (28.2.8).

(ii) Verify (28.3.5).

(iii) Derive (28.3.6) and (28.3.9).

(iv) Explain why the current distribution on a short dipole can be approximated by a triangle.

(v) Explain why retardation effect is ignored in such calculations in (28.3.8).

Chapter 29

Radiation Fields, Directive Gain, Effective Aperture

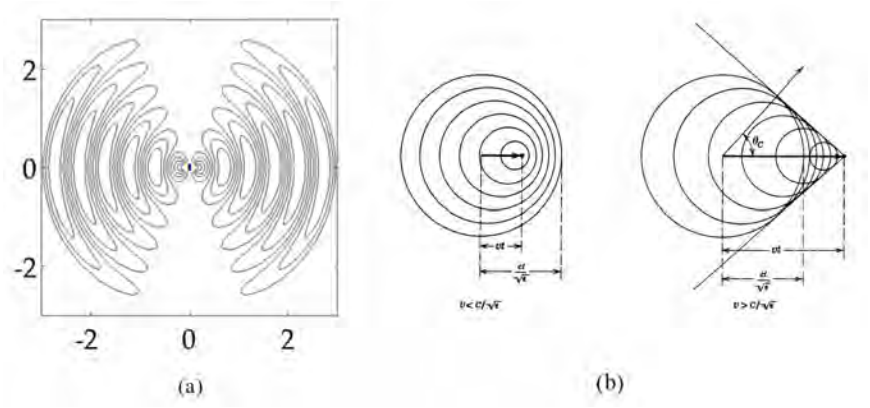


Figure 29.1: (a) Electric field around a time-oscillating dipole (courtesy of physics stack exchange). (b) Equipotential lines around a moving charge that gives rise to Cherenkov radiation (courtesy of J.D. Jackson [49]). We will not study Cherenkov (Cerenkov) radiation in this course, but it is written up in [49] and [34]. Its discovery and its explanation led to a Nobel Prize.

The reason why charges radiate is because they move or accelerate. In the case of a dipole antenna, the charges move back and forth between the two poles of the antenna. Near to the dipole source, quasi-static physics prevails, and the field resembles that of a static dipole. If the dipole is flipping sign constantly due to the change in the direction of the current flow, the field would also have to flip sign constantly. But electromagnetic wave travels with a finite velocity. The field from the

source ultimately cannot keep up with the sign change of the source field: it has to be 'torn' away from the source field and radiate. Another interesting radiation is the Cherenkov (also spelled Cerenkov) radiation. It is due to a charge moving faster than the velocity of light. As an electron cannot move faster than the speed of light in a vacuum, this can only happen in the material media or plasma, where the velocity of the electron can be faster than the group velocity of wave in the medium. Ultimately, the electric field from the particle is 'torn' off from the charge and radiate. These two kinds of radiation are shown in the Figure 29.1.

We have shown how to connect the vector and scalar potentials to the sources $\mathbf{J}$ and ϱ of an electromagnetic system in our previous lectures. This is a very important connection: it implies that once we know the sources, we know how to find the fields $\mathbf{A}$ and Φ, and then $\mathbf{E}$ and $\mathbf{H}$. But the relation between the fields and the sources are in general rather complex. In this lecture, we will simplify this relation by making a radiation field or far-field approximation. To this end, we assume that the point where the field is observed is very far from the source location in terms of wavelength. This approximation is very useful for understanding the physics of the radiation field from a source such as an antenna. It is also important for understanding the far field of an optical system. As shall be shown, this radiation field carries the energy generated by the sources to infinity.

29.1 Radiation Fields or Far-Field Approximation

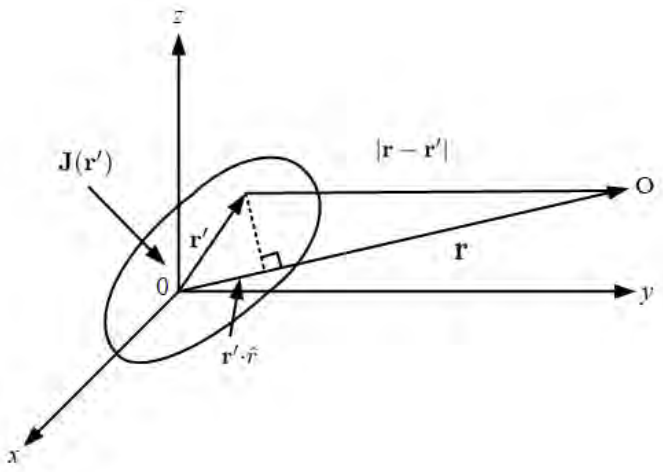


Figure 29.2: The relation of the observation point located at $\mathbf{r}$ to the source location at $\mathbf{r}'$. The distance of the observation point $\mathbf{r}$ to the source location $\mathbf{r}'$ is $|\mathbf{r} - \mathbf{r}'|$.

In the previous lecture, we have derived the relation of the vector and scalar potentials to the sources $\mathbf{J}$ and ϱ as shown in (27.2.29) and (27.2.30).[1] They are given by

$$\mathbf{A}(\mathbf{r}) = \mu \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{29.1.1}$$

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon} \iiint_V d\mathbf{r}' \varrho(\mathbf{r}') \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{29.1.2}$$

where $\beta = \omega\sqrt{\mu\varepsilon} = \omega/c$ is the wavenumber. The integrals in (29.1.1) and (29.1.2) are normally untenable, but when the observation point is far from the source, approximation to the integrals can be made giving them a nice physical interpretation.

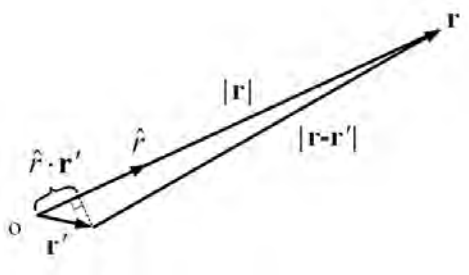


Figure 29.3: The relation between $|\mathbf{r}|$ and $|\mathbf{r}-\mathbf{r}'|$ using the parallax method, or that $|\mathbf{r}-\mathbf{r}'| \approx |\mathbf{r}| - \mathbf{r}' \cdot \hat{r}$. It is assumed that $\mathbf{r}$ is almost parallel to $\mathbf{r}-\mathbf{r}'$.

29.1.1 Far-Field Approximation

When $|\mathbf{r}| \gg |\mathbf{r}'|$, then $|\mathbf{r}-\mathbf{r}'| \approx r - \mathbf{r}' \cdot \hat{r}$, where $r = |\mathbf{r}|$. This approximation can be shown algebraically[2] or by geometrical argument as shown in Figure 29.3. Thus (29.1.1) above becomes

$$\mathbf{A}(\mathbf{r}) \approx \frac{\mu}{4\pi} \iiint_V d\mathbf{r}' \frac{\mathbf{J}(\mathbf{r}')}{r - \mathbf{r}' \cdot \hat{r}} e^{-j\beta r + j\beta \mathbf{r}' \cdot \hat{r}} \approx \frac{\mu e^{-j\beta r}}{4\pi r} \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') e^{j\beta \mathbf{r}' \cdot \hat{r}} \tag{29.1.3}$$

In the above, $\mathbf{r}' \cdot \hat{r}$ is small compared to r. Hence, we have made use of that $1/(1-\Delta) \approx 1$ when Δ is small, so that $1/(r - \mathbf{r}' \cdot \hat{r})$ can be approximate by $1/r$. Also, we assume that the frequency is sufficiently high such that $\beta \mathbf{r}' \cdot \hat{r}$ is not necessarily small. Thus, $e^{j\beta \mathbf{r}' \cdot \hat{r}} \neq 1$, unless $\beta \mathbf{r}' \cdot \hat{r} \ll 1$. Hence, we keep the exponential term in (29.1.3) but simplify the denominator to arrive at the last expression above.

If we let $\boldsymbol{\beta} = \beta \hat{r}$, which is the $\boldsymbol{\beta}$ vector (or $\mathbf{k}$ vector in optics), and let $\mathbf{r}' = \hat{x}x' + \hat{y}y' + \hat{z}z'$, then we can convert the factor in the integrand into one that looks like a plane wave or 3D Fourier

[1] This topic is found in many standard textbooks in electromagnetics [57, 51, 34]. They are also found in lecture notes [47, 196].

[2] To show this algebraically, we let $|\mathbf{r}-\mathbf{r}'|^2 = (\mathbf{r}-\mathbf{r}') \cdot (\mathbf{r}-\mathbf{r}') = |\mathbf{r}|^2 + |\mathbf{r}'|^2 - 2\mathbf{r} \cdot \mathbf{r}' \cong |\mathbf{r}|^2 - 2\mathbf{r} \cdot \mathbf{r}'$. By taking the square root of this approximation, and using the algebraic approximation that $(1-x)^{0.5} \cong 1 - 0.5x$, we arrive at the given approximation.

transform. Namely,

$$e^{j\beta \mathbf{r}'\cdot\hat{r}} = e^{j\boldsymbol{\beta}\cdot\mathbf{r}'} = e^{j\beta_x x' + j\beta_y y' + j\beta_z z'} \tag{29.1.4}$$

The above is the expression for a plane wave propagating in the $\boldsymbol{\beta} = \beta\hat{r}$ direction or $\hat{r}$ direction. Therefore (29.1.3) resembles a 3D Fourier transform integral,[3] namely, the above integral becomes

$$\mathbf{A}(\mathbf{r}) \approx \frac{\mu e^{-j\beta r}}{4\pi r} \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') e^{j\boldsymbol{\beta}\cdot\mathbf{r}'} \tag{29.1.5}$$

and (29.1.5) can be rewritten as

$$\mathbf{A}(\mathbf{r}) \cong \frac{\mu e^{-j\beta r}}{4\pi r} \mathbf{F}(\boldsymbol{\beta}) \tag{29.1.6}$$

where

$$\mathbf{F}(\boldsymbol{\beta}) = \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') e^{j\boldsymbol{\beta}\cdot\mathbf{r}'} \tag{29.1.7}$$

It is the 3D Fourier transform of $\mathbf{J}(\mathbf{r}')$ with the Fourier transform variable $\boldsymbol{\beta} = \hat{r}\beta$. In a word, the Fourier data is restricted to be on a sphere surface with radius β, which is not usual, or that $|\boldsymbol{\beta}|^2 = {\beta_x}^2 + {\beta_y}^2 + {\beta_z}^2 = \beta^2$ which is a constant for a fixed frequency. In other words, the length of the vector $\boldsymbol{\beta}$ is fixed to be β, whereas in a usual 3D Fourier transform, β_x, β_y, and β_z are independent variables. Or the value of ${\beta_x}^2 + {\beta_y}^2 + {\beta_z}^2$ ranges from zero to infinity, but in (29.1.7), it is different.

The above is the 3D "Fourier transform" of the current source $\mathbf{J}(\mathbf{r}')$ with Fourier variables, β_x, β_y, β_z restricted to lying on a sphere of radius β and $\boldsymbol{\beta} = \beta\hat{r}$. This spherical surface in the Fourier space is also called the *Ewald sphere*.

29.1.2 Locally Plane Wave Approximation

We can write $\hat{r}$ or $\boldsymbol{\beta}$ in terms of direction cosines in spherical coordinates or that

$$\hat{r} = \hat{x}\cos\phi\sin\theta + \hat{y}\sin\phi\sin\theta + \hat{z}\cos\theta \tag{29.1.8}$$

Hence,

$$\mathbf{F}(\boldsymbol{\beta}) = \mathbf{F}(\beta\hat{r}) = \mathbf{F}(\beta, \theta, \phi) \tag{29.1.9}$$

It is not truly a 3D function in space, since β, the length of the vector $\boldsymbol{\beta}$, is fixed; thus it is a function of θ and ϕ space variables. Or it is a 3D Fourier transform with data restricted on a spherical surface.

In (29.1.6), when $r \gg \mathbf{r}' \cdot \hat{r}$, and when the frequency is high or β is large, $e^{-j\beta r}$ is now a rapidly varying function of r while, $\mathbf{F}(\boldsymbol{\beta})$ is only a slowly varying function of $\hat{r}$ or of θ and of ϕ, the observation angles. In other words, the prefactor in (29.1.6), $\exp(-j\beta r)/r$, resembles a spherical wave which is modulated by a slowly varying function of θ and ϕ, or $\mathbf{F}(\boldsymbol{\beta})$.

[3] Except that the vector $\boldsymbol{\beta}$ is of fixed length.

Therefore, in the far field, the wave radiated by a finite source resembles a spherical wave. Moreover, a spherical wave resembles a plane wave when one is sufficiently far from the source such that $\beta r \gg 1$, or $2\pi r/\lambda \gg 1$. This happens when r is many wavelengths away from the source. This is obviated if we write $e^{-j\beta r} = e^{-j\boldsymbol{\beta}\cdot\mathbf{r}}$ where $\boldsymbol{\beta} = \hat{r}\beta$ and $\mathbf{r} = \hat{r}r$ so that a spherical wave resembles a plane wave locally. This phenomenon is shown in Figure 29.4 and Figure 29.5

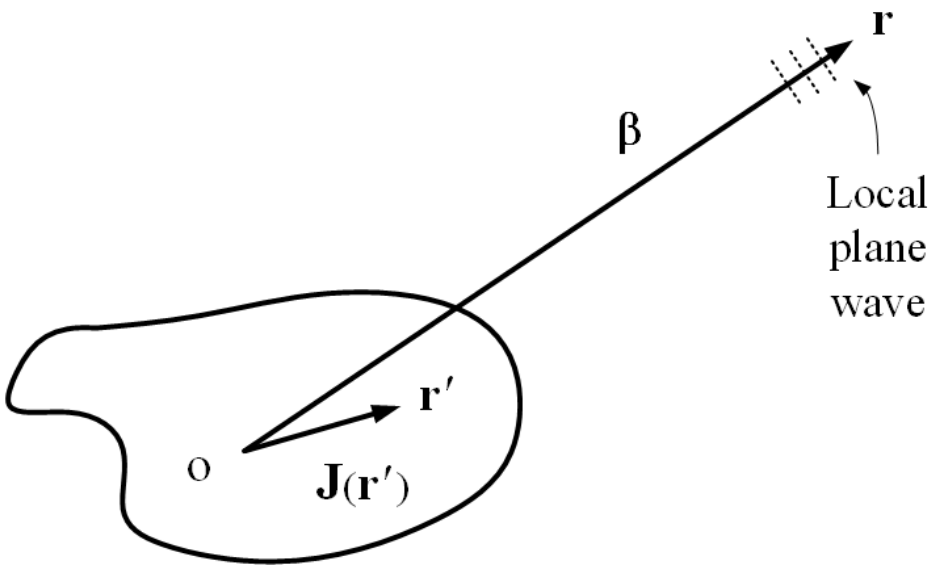


Figure 29.4: A finite source radiates a field that resembles a spherical wave far from the source. In the vicinity of the observation point $\mathbf{r}$, when β is large, the field is strongly dependent on r via $\exp(-j\beta r)$ but weakly dependent on $\boldsymbol{\beta}$ ($\boldsymbol{\beta}$ hardly changes direction in the vicinity of the observation point $\mathbf{r}$). Hence, the field becomes locally like a plane wave in the far field.

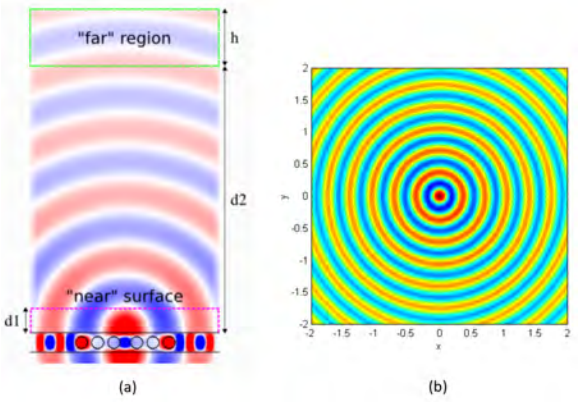


Figure 29.5: (a) A leaky hole in a waveguide leaks a spherical (courtesy of MEEP, MIT). (b) A point source radiates a spherical wave (courtesy of ME513, Purdue Engineering). Most of these simulations are done with FDTD (finite-difference time-domain) method that we will learn later in the course. When the wavelength is short, or the frequency high, a spherical wave looks locally a plane wave. This is similar to the notion that as humans, who are small, think that the earth is flat around us. Up to this day, some people still believe that the earth is flat:)

Then, it is clear that with the local plane-wave approximation, we can let $\nabla \to -j\boldsymbol{\beta} = -j\beta\hat{r}$, and with this approximation, we have

$$\mathbf{H} = \frac{1}{\mu}\nabla \times \mathbf{A} \approx -j\frac{\beta}{\mu}\hat{r} \times (\hat{\theta}A_\theta + \hat{\phi}A_\phi) = j\frac{\beta}{\mu}(\hat{\theta}A_\phi - \hat{\phi}A_\theta) \tag{29.1.10}$$

Similarly [47, 196],

$$\mathbf{E} = \frac{1}{j\omega\varepsilon}\nabla \times \mathbf{H} \cong -j\frac{\beta}{\omega\epsilon}\hat{r} \times \mathbf{H} \cong -j\omega(\hat{\theta}A_\theta + \hat{\phi}A_\phi) \tag{29.1.11}$$

Notice that $\boldsymbol{\beta} = \beta\hat{r}$, the direction of propagation of the local plane wave, is orthogonal to $\mathbf{E}$ and $\mathbf{H}$ in the far field. This is a property of a plane wave since the wave is locally a plane wave.

Moreover, there are more than one way to derive the electric field $\mathbf{E}$. For instance, using (29.1.10) for the magnetic field, the electric field can also be written as

$$\mathbf{E} = \frac{1}{j\omega\mu\varepsilon}\nabla \times (\nabla \times \mathbf{A}) \tag{29.1.12}$$

Using the BAC-CAB formula for the double-curl operator, plus that $\nabla \simeq -j\boldsymbol{\beta}$, for a plane wave, the above can be rewritten as

$$\mathbf{E} = \frac{1}{j\omega\mu\varepsilon}\left(\nabla\nabla \cdot \mathbf{A} - \nabla^2\mathbf{A}\right) \cong \frac{1}{j\omega\mu\varepsilon}\left(-\boldsymbol{\beta}\boldsymbol{\beta} + \beta^2\bar{\mathbf{I}}\right) \cdot \mathbf{A} \tag{29.1.13}$$

where we have used that $\nabla^2\mathbf{A} = -\beta^2\mathbf{A} = -\beta^2\bar{\mathbf{I}}\cdot\mathbf{A}$, where $\bar{\mathbf{I}}$ is the identity dyad.[4]

Alternatively, we bring $\beta^2 = \omega^2\mu\epsilon$ out of the parenthesis, and rewrite the above as

$$\mathbf{E} \cong -j\omega\left(-\hat{\beta}\hat{\beta} + \bar{\mathbf{I}}\right)\cdot\mathbf{A} = -j\omega\left(-\hat{r}\hat{r} + \bar{\mathbf{I}}\right)\cdot\mathbf{A} \tag{29.1.14}$$

Since we can express the identity dyad $\bar{\mathbf{I}} = \hat{r}\hat{r} + \hat{\theta}\hat{\theta} + \hat{\phi}\hat{\phi}$ in spherical coordinates,[5] then the above becomes

$$\mathbf{E} \cong -j\omega\left(\hat{\theta}\hat{\theta} + \hat{\phi}\hat{\phi}\right)\cdot\mathbf{A} = -j\omega(\hat{\theta}A_\theta + \hat{\phi}A_\phi) \tag{29.1.15}$$

which is the same as previously derived. It also shows that the electric field is transverse to the $\boldsymbol{\beta}$ vector.[6]

Furthermore, it can be shown that in the far field, using the locally plane-wave approximation,

$$|\mathbf{E}|/|\mathbf{H}| \approx \eta \tag{29.1.16}$$

where η is the intrinsic impedance of free space, which is a property of a plane wave. Moreover, one can show that the time average Poynting's vector, or the power density flow, in the far field is

$$\langle\mathbf{S}\rangle = \frac{1}{2}\Re e\left(\mathbf{E}\times\mathbf{H}^*\right) \approx \frac{1}{2\eta}|\mathbf{E}|^2\hat{r} = \langle S_r\rangle\hat{r} \tag{29.1.17}$$

which again, resembles also the property of a plane wave.[7] In a word, the radiated field is a spherical wave, the Poynting's vector is radial. Therefore,

$$\langle\mathbf{S}\rangle = \hat{r}\langle S_r(\theta,\phi)\rangle, \quad \text{where} \quad \langle S_r(\theta,\phi)\rangle = \frac{1}{2\eta}|\mathbf{E}|^2 \tag{29.1.18}$$

and $\langle S_r\rangle$ is the time-average radial power density. The plot of $|\mathbf{E}(\theta,\phi)|$ is termed the far-field pattern or the radiation pattern of an antenna or the source, while the plot of $|\mathbf{E}(\theta,\phi)|^2$ is its far-field power pattern.

29.1.3 Directive Gain Pattern

A directive gain pattern that characterizes the radiation pattern of a general source or an antenna can be defined. Once the far-field radiation power pattern or the radial power density $\langle S_r\rangle$ is known, the total power radiated by the antenna in the far field is found by integrating over all angles, viz.,

$$P_T = \int d\Omega\langle S_r(\Omega)\rangle\int_0^\pi\int_0^{2\pi} r^2\sin\theta d\theta d\phi\langle S_r(\theta,\phi)\rangle \tag{29.1.19}$$

[4]Note that $\nabla\cdot\mathbf{A} \neq 0$ here.

[5]Easily verified by a sanity check.

[6]We can also arrive at the above by letting $\mathbf{E} = -j\omega\mathbf{A} - \nabla\Phi$, and using the appropriate formula for the scalar potential. There is more than one road that lead to Rome!

[7]To avoid confusion, we will use $\mathbf{S}$ to denote instantaneous Poynting's vector and $\underline{\mathbf{S}}$ to denote complex Poynting's vector (see 10.3.1).

In free space, the above evaluates to a constant independent of r due to energy conservation.

Now assume that this same antenna is radiating isotropically in all directions; then the average power density of this fictitious isotropic radiator as $r \to \infty$ is

$$\langle S_{\text{av}} \rangle = \frac{P_T}{4\pi r^2} \tag{29.1.20}$$

A dimensionless directive gain pattern can be defined such that [34, 196]

$$G(\theta, \phi) = \frac{\langle S_r(\theta, \phi) \rangle}{\langle S_{\text{av}} \rangle} = \frac{4\pi r^2 \langle S_r(\theta, \phi) \rangle}{P_T} \tag{29.1.21}$$

Now we can use the Hertzian dipole to illustrate this concept. Using P_T from (28.3.5), and $\langle S_r \rangle$ from (28.3.1), we arrive at

$$G(\theta, \phi) = \frac{\frac{\eta_0}{2} \left(\frac{\beta I l}{4\pi r} \right)^2 \sin^2 \theta}{\frac{1}{4\pi r^2} \frac{4}{3} \eta_0 \pi \left(\frac{\beta I l}{4\pi} \right)^2} = \frac{3}{2} \sin^2 \theta \tag{29.1.22}$$

This directive gain pattern is a measure of the radiation power pattern of the antenna or source compared to when it radiates isotropically. The above function is independent of r in the far field since $S_r \sim 1/r^2$ in the far field but it is a function of (θ, ϕ).

The directivity of an antenna $D = \max(G(\theta, \phi))$ is the maximum value of the directive gain. This is 1.5 for the Hertzian dipole. It is to be noted that by its mere definition,

$$\int d\Omega G(\theta, \phi) = 4\pi \tag{29.1.23}$$

where $\int d\Omega = \int_0^{2\pi} \int_0^{\pi} \sin\theta d\theta d\phi$. It is seen that since the directive gain pattern is normalized, when the radiation power is directed to the main lobe of the antenna, the corresponding side lobes and back lobes will be diminished. This follows from energy conservation.

29.2 Effective Aperture and Directive Gain

An antenna also has an effective area or aperture A_e, such that if a plane wave carrying power density denoted by $\langle S_{\text{inc}} \rangle$ impinges on the antenna, then the power received by the antenna, P_{received} is given by

$$P_{\text{received}} = \langle S_{\text{inc}} \rangle A_e \tag{29.2.1}$$

Here, the transmit antenna and the receive antenna are in the far field of each other. Hence, we can approximate the field from the transmit antenna to be a plane wave when it reaches the receive antenna. The receive antenna, if it is a good antenna, will siphon off some energy from the plane wave to power the local electronics within the antenna. Therefore, the antenna absorbs some energy to generate a voltage at the receiver load to deliver the power received by the antenna. This receiving system can be linearly modeled and the received power is linearly proportional to

the incident power density $\langle S_{\text{inc}} \rangle$ with the proportionality constant which is the effective aperture of the antenna.[8]

A wonderful relationship exists between the directive gain pattern $G(\theta, \phi)$ and the effective aperture, namely that

$$A_e(\theta, \phi) = \frac{\lambda^2}{4\pi} G(\theta, \phi) \tag{29.2.2}$$

Therefore, the effective aperture of an antenna is also direction dependent. Different plane waves incident on the antenna at different angles will see different effective apertures. The above implies that the radiation property of an antenna is related to its receiving property. (This is a beautiful consequence of reciprocity theorem! The constant of proportionality, $\lambda^2/(4\pi)$ is a universal constant that is valid for all antennas satisfying reciprocity theorem. The derivation of this constant for a Hertzian dipole is given in Kong [34], or using blackbody radiation law [196, 197].)

The directivity and the effective aperture can be enhanced by designing antennas with different gain patterns. When the radiative power of the antenna can be directed to be in a certain direction, then the directive gain and the effective aperture (for that given direction) of the antenna is improved. This is shown in Figure 29.6. Such focussing of the radiation fields of the antenna can be achieved using reflector antennas or array antennas. Array antennas, as shall be shown, work by constructive and destructive wave field of the antenna.

Being able to do point-to-point communications at high data rate is an important modern application of antenna array. Figure 29.7 shows the gain pattern of a sophisticated antenna array design for 5G applications. Now, 5G antennas operating in the high band usually range in frequencies from 24 GHz to 50 GHz, putting them in the millimeter wave range, given the antennas much higher data throughput or bandwidth.

[8] This concept can be extended to effective cross section or aperture of an atom that has the dimension of area.

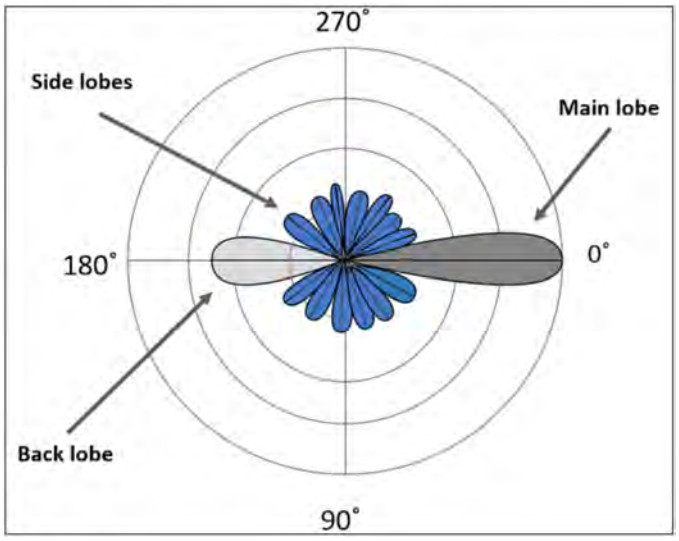


Figure 29.6: The directive gain pattern of an array antenna. The directivity is increased by constructive interference among the elements of the array antenna (courtesy of Wikepedia).

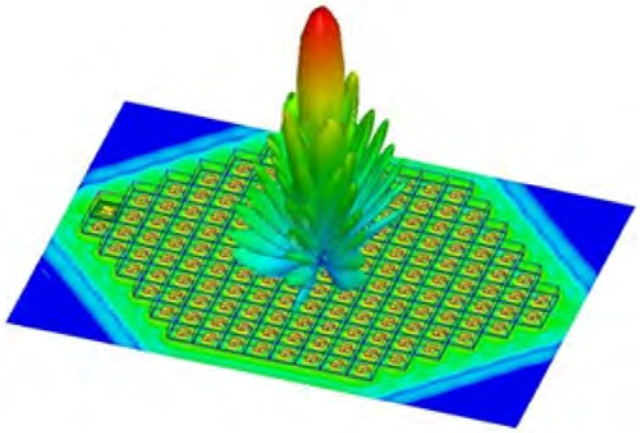


Figure 29.7: The directive gain pattern of a sophisticated array antenna for 5G applications. Presently, 5G antennas work from 24 GHz to 50 GHz which are in the millimeter range. (courtesy of Ozeninc.com).

29.2.1 The Electromagnetic Spectrum

Now that we have learnt more about the physics of electromagnetic waves, it is meaningful to discuss the electromagnetic spectrum. The electromagnetic field has been used from very low frequencies to very high frequencies. At very low frequencies, ultra-low frequency (ULF) less than 3 Hz, extremely-low frequency (ELF) 3-3000 Hz, very low frequency (VLF) 3 KHz to 30 KHz have been used to probe the earth surface, and submarine communication because of their deeper penetration depths. This is because in a saline solution, which is primarily a conductive medium, has larger skin depth the lower the frequency. Remember that the skin depth is given by the formula $\delta = \sqrt{2/(\omega\mu\sigma)}$!

The AM radio stations, operate in the several hundreds KHz, have wavelengths of several hundreds meters. FM radio are in the 100 MHz range, while TV stations operate in the several hundreds MHz. Microwaves have wavelengths of order of cm, and infra-red light ranges from 1000 μm to 1 μm. A perfect electric conductor (PEC) has zero skin depth, but when the skin depth is much smaller than the wavelength and the size of the object, then one can approximate a conductive medium as a PEC, e.g., copper at microwave frequencies.

The visible spectrum ranges from 700 nm to 400 nm. There is no PEC at these optical frequencies. As we have learnt from the Drude-Lorentz-Sommerfeld model, when the frequency is high, the inertial term in (8.3.14) in the Chapter 8, which is proportional to $m_e d^2x/dt^2$ is important. When this term dominates, the medium resembles a plasma medium.

Ultra-violet (UV) light ranges from 400 nm to 100 nm, while X-ray is generally below 100 nm to 1 nm. Gamma ray is generally below 1 nm. UV light of 193 nm and EUV light of 13.5 nm are now used for nano-lithography. X-ray is important for imaging, while gamma ray is used for some medical applications. The lights above UV are ionizing, and thus generally harmful to the human body. At high frequencies, it is expedient to think of electromagnetic waves as consisting of stream of particles which are photons.

The new frontier of the electromagnetic spectrum is in the terahertz range, from 0.1 THz to 10 THz. The dearth of THz sources has made the technology in this area in its infancy. THz signals can be generated from frequency doubling from lower semiconductor devices, nonlinear mixing of optical signals, or the use of quantum cascade lasers.

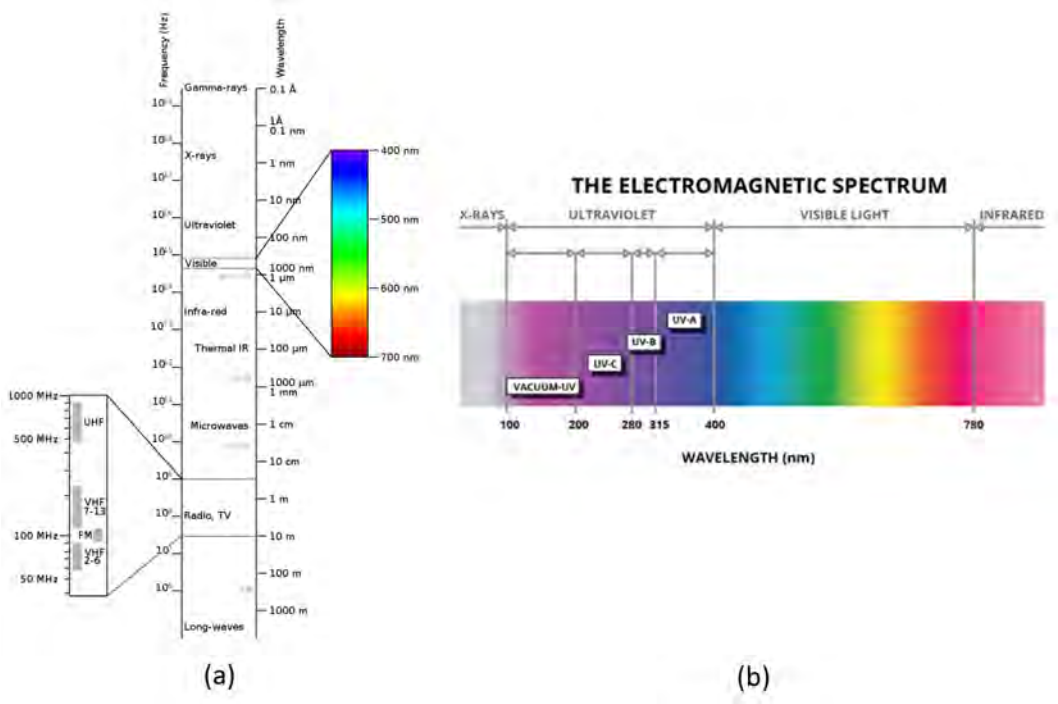


Figure 29.8: (a) The electromagnetics spectrum from static to gamma ray. (b) The electromagnetics spectrum with emphasis around the visible/optical spectrum. EUV with wavelength of 13.5 nm is used for nanolithography now, while ELF with frequency of around 3 Hz are used for submarine communications (courtesy of Wikepedia and Bestlight.io).

Exercises for Lecture 29

Problem 29-1:

(i) Show that in the far field, the ratio of $|\mathbf{E}|/|\mathbf{H}| = \eta$. Give the physical reason for this.

(ii) In this lecture, it was shown that

$$\mathbf{A}(\mathbf{r}) \cong \frac{\mu e^{-j\beta r}}{4\pi r}\mathbf{F}(\boldsymbol{\beta})$$

Verify the above expression, and give the expression for $\mathbf{F}(\boldsymbol{\beta})$.

(iii) Explain why when we are in the far field such that $\beta r \gg 1$, in the neighborhood of an observation around $\mathbf{r}$, the variation of the field is dominated by $e^{-j\beta r}$, and hence, why we can make a local plane wave approximation. Is it necessary that the radius of curvature of the wavefront be much larger than the wavelength for this approximation? Explain why and why not?

(iv) For an antenna, give the physical meaning for the directive gain pattern $G(\theta, \phi)$ defined in (29.1.21) and the effective aperture A_e defined in (29.2.1).

Chapter 30

Array Antennas, Fresnel Zone, Rayleigh Distance

Our world is beset with the dizzying impact of wireless communication. It has greatly impacted our lives.[1] Wireless communication is impossible without using antennas. Hence it is important to design these communication systems with the proper antennas so that they operate with utmost efficiency and sensitivity. We have seen that a simple Hertzian dipole has low directivity in Section 29.1.3. The radiation pattern looks like that of a donut, and the directivity of the Hertzian dipole antenna is 1.5. Hence, for point-to-point communications, much power is wasted. However, the directivity of antennas can be improved if a group or array of dipoles can work cooperatively together by using constructive and destructive interferences. They can be made to constructively interfere in the desired direction, and destructively interfere in other directions to enhance the directivity of the array of antennas. Since the far-field approximation of the radiation field can be made, and the relationship between the far field and the source is a Fourier transform relationship,[2] clever engineering can be done by borrowing knowledge from the signal processing area. After understanding the far-field physics, one can also understand many optical phenomena, such as how a laser pointer works. Many textbooks have been written about array antennas some of which are in the reference list [198, 199].

We like to emphasize that the world of electromagnetics morphs from circuit physics to wave physics when the wavelength becomes small and the structure size is of the order of wavelength. If the frequency increases such that the wavelength becomes much smaller than the feature sizes, then ray physics prevails, and electromagnetics waves behave like particles. We will study the wave physics nature of electromagnetic fields in this chapter. Later, we will address the ray nature of electromagnetic fields.

[1] I cannot imagine a day that I do not receive messages from my relatives and friends in Malaysia and Singapore. It has made the world look smaller, and more transparent.

[2] It actually is a Fourier transform relationship on an Ewald sphere.

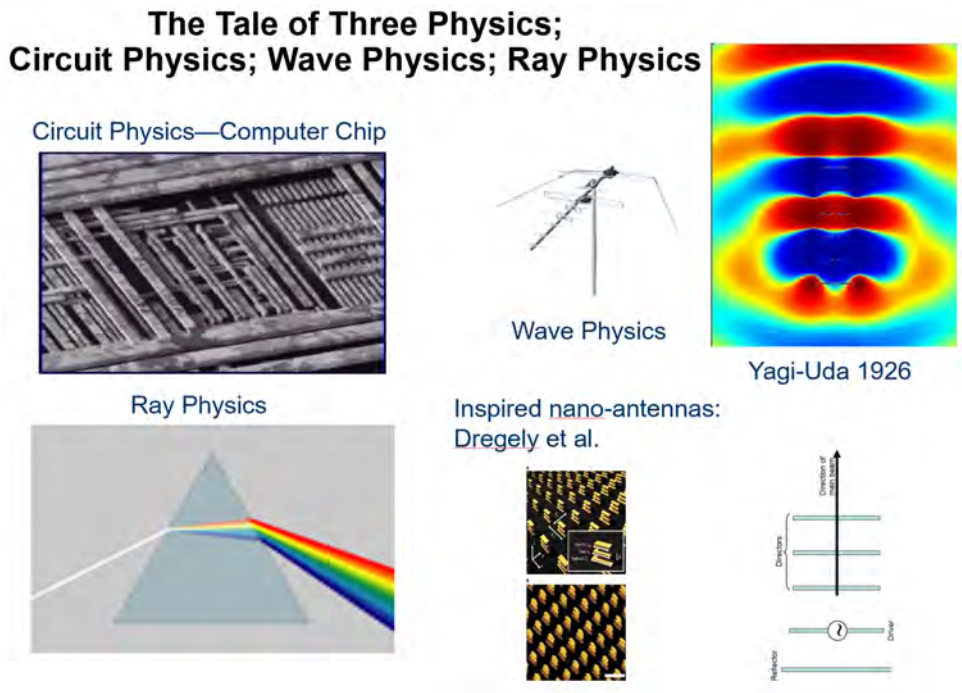


Figure 30.1: The world of electromagnetics morphs from circuit physics to wave physics, and then ray physics. The plethora of physics that emerges in electromagnetics is fascinating. Different mathematical tools are devised to solve electromagnetics problems in these regimes. At low frequencies, KCL, KVL, and potential theory can be use to describe the working of most devices. At mid-frequencies, full Maxwell's equations have to be solved to capture the wave physics correctly. At high frequencies, electromagnetic waves behaves like particles and rays. Ray tracing algorithms are often used to capture their wave physics. What is missing here is the quantum nature of electromagnetic fields that will be discussed at the end of this course.

30.1 Array Pattern–Unit Pattern and Array Factor

The fact that the far field of an antenna is proportional to the Fourier transform of the current on the antenna can be exploited for further insight in many areas of wave physics. This fact is important, and hence, (29.1.3) is reproduced below:

$$\mathbf{A}(\mathbf{r}) \approx \frac{\mu e^{-j\beta r}}{4\pi r} \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') e^{j\boldsymbol{\beta}\cdot\mathbf{r}'} \cong \frac{\mu e^{-j\beta r}}{4\pi r} \mathbf{F}(\boldsymbol{\beta}) \tag{30.1.1}$$

In the above,

$$\mathbf{F}(\boldsymbol{\beta}) = \iiint_V d\mathbf{r}'\mathbf{J}(\mathbf{r}')e^{j\boldsymbol{\beta}\cdot\mathbf{r}'} \tag{30.1.2}$$

The above is the Fourier transform of the radiation current restricted to the Ewald sphere.

In an antenna array, we make cookie-cutter replicas of a primary antenna, and design them to radiate in unison. By appropriately adjusting their mutual phase relations and amplitudes, we can engineer and produce a rich diversity of radiation patterns by constructive and destructive interference. For simplicity, we begin with two antenna elements which are the replica of each other, but they have unequal amplitudes and phases. Mathematically, this can be expressed as

$$\mathbf{J}(\mathbf{r}') = \mathbf{J}_u(\mathbf{r}') + B\mathbf{J}_u(\mathbf{r}' - \mathbf{r}_T) \tag{30.1.3}$$

where B is a complex amplitude and hence, the two currents are not equi-phase. Moreover, the second current is being translated to a new location $\mathbf{r}_T$.

Using the Fourier transform property that if

$$\mathbf{J}_u(\mathbf{r}') \quad \text{Fourier Transforms To:} \quad \mathbf{F}_u(\beta) \tag{30.1.4}$$

then

$$\mathbf{J}_u(\mathbf{r}' - \mathbf{r}_T) \quad \text{Fourier Transforms To:} \quad \mathbf{F}_u(\boldsymbol{\beta})e^{-j\boldsymbol{\beta}\cdot\mathbf{r}_T} \tag{30.1.5}$$

Consequently, we can Fourier transform (30.1.3) to get

$$\mathbf{F}(\boldsymbol{\beta}) = \mathbf{F}_u(\boldsymbol{\beta})(1 + Be^{-j\boldsymbol{\beta}\cdot\mathbf{r}_T}) \tag{30.1.6}$$

And then, $\mathbf{A}(\mathbf{r})$ in (30.1.1), which is a Fourier transform integral, becomes

$$\mathbf{A}(\mathbf{r}) \cong \underbrace{\frac{\mu e^{-j\beta r}}{4\pi r}\mathbf{F}_u(\boldsymbol{\beta})}_{\mathbf{A}_u(\mathbf{r})} \underbrace{(1 + Be^{-j\boldsymbol{\beta}\cdot\mathbf{r}_T})}_{A.F.} \tag{30.1.7}$$

where

$$\mathbf{A}_u(\mathbf{r}) = \frac{\mu e^{-j\beta r}}{4\pi r}\mathbf{F}_u(\boldsymbol{\beta}) \tag{30.1.8}$$

In the above, $\mathbf{A}_u(\mathbf{r})$ is the field due to a unit element of the array, and $A.F.$ is the array factor given as

$$A.F. = 1 + Be^{-j\boldsymbol{\beta}\cdot\mathbf{r}_T} \tag{30.1.9}$$

Using that in the far electric field,

$$\mathbf{E} \cong -j\omega(\hat{\theta}A_\theta + \hat{\phi}A_\phi) = -j\omega\left(\hat{\theta}\hat{\theta} + \hat{\phi}\hat{\phi}\right)\cdot\mathbf{A} = -j\omega\left(\hat{\theta}\hat{\theta} + \hat{\phi}\hat{\phi}\right)\cdot\mathbf{A}_u A.F. \tag{30.1.10}$$

the far electric field due to a unit element of the array is thus

$$\mathbf{E}_u \cong -j\omega\left(\hat{\theta}\hat{\theta} + \hat{\phi}\hat{\phi}\right)\cdot\mathbf{A}_u \tag{30.1.11}$$

Then, using (34.3.10) and (34.3.11), the far field pattern can be written as

$$|\mathbf{E}| = |\mathbf{E}_u| A.F. \tag{30.1.12}$$

Therefore, the far field pattern is the product of the unit pattern with the array factor $A.F.$. The array factor above is for a two-element array. But $A.F.$ can be generalized to that of an N-element array to give

$$A.F. = 1 + B_1 e^{-j\boldsymbol{\beta}\cdot\mathbf{r}_{T1}} + B_2 e^{-j\boldsymbol{\beta}\cdot\mathbf{r}_{T2}} + \cdots = \sum_{i=0}^{N} B_i e^{-j\boldsymbol{\beta}\cdot\mathbf{r}_{Ti}} \tag{30.1.13}$$

where we take $B_0 = 1$.

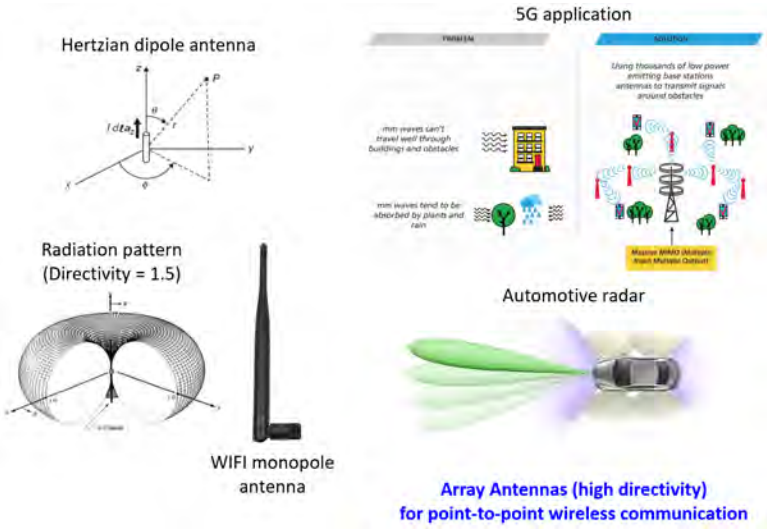


Figure 30.2: The left figures show the low directivity of a Hertzian dipole. The right shows applications where array antennas are needed to improve directivity and better communication links of a 5G communication system which can operate as high as 54 GHz. The bottom right figure shows a collision avoidance system (CAS) of a car. The CAS of cars typically operate around 77 GHz so that the electromagnetics signal can penetrate through fog and rain (assembled by D.Y. Na).

30.2 Linear Array of Dipole Antennas

Antenna array can be designed so that the constructive and destructive interference in the far field can be used to steer the direction of radiation of the antenna, or the far-field radiation pattern of an antenna array. This is because the far field of a source is related to the source by a Fourier transform relationship as shown in the previous lecture (see Subsection 29.1.1). The relative phases of the array elements can be changed slowly in time with respect to the operating frequency so

that the beam of an array antenna can be steered in real time. This has important applications in, for example, air-traffic control. It is to be noted that if the current sources are impressed current sources, they can be regarded as the input to Maxwell's equations. Then the fields are the output of the system, and we are dealing with a linear time-invariant system here whereby linear system theory can be used. For instance, we can use Fourier transform to analyze the problem in the frequency domain. The time domain response then can be obtained by inverse Fourier transform. This is provided that the current sources are impressed and time-invariant. They are unaffected by the fields that they radiate.

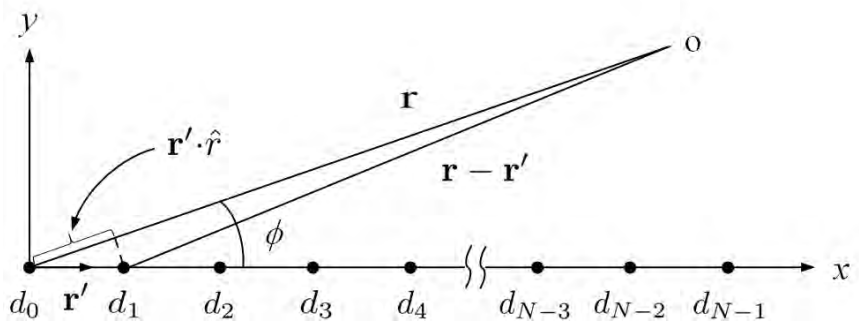


Figure 30.3: Schematic of a dipole array where the dipoles are aligned on the x axis, and we observe the field on the xy plane. To simplify the math, the far-field approximation can be used to find its far field or radiation field.

To gain physical insight into how constructive and destructive interference works for an antenna array, we assume a simple linear dipole array as shown in Figure 30.3. First, without loss of generality and for simplicity to elucidate the physics, we assume that this is a linear array of point Hertzian dipoles aligned on the x axis. The current can then be described mathematically as follows:

$$\begin{aligned}\mathbf{J}(\mathbf{r}') = \hat{z}Il[&A_0\delta(x') + A_1\delta(x'-d_1) + A_2\delta(x'-d_2) + \cdots \\ &+ A_{N-1}\delta(x'-d_{N-1})]\delta(y')\delta(z')\end{aligned} \tag{30.2.1}$$

Again for simplicity, all the dipoles are pointing in the z axis. The far field can be found using the approximate formula derived in the previous lecture, viz., (29.1.3) reproduced below:

$$\mathbf{A}(\mathbf{r}) \approx \frac{\mu e^{-j\beta r}}{4\pi r} \iiint_V d\mathbf{r}'\mathbf{J}(\mathbf{r}')e^{j\boldsymbol{\beta}\cdot\mathbf{r}'} \tag{30.2.2}$$

To reiterate, the above implies that the far field is related to the Fourier transform of the current source $\mathbf{J}(\mathbf{r}')$.[3]

[3]Again, this is an unusual kind of Fourier transform where the length of the $\boldsymbol{\beta}$ vector is fixed. Nevertheless, one can gain a lot of insight by using our knowledge of Fourier transform to understand a number of physical phenomena. For instance, it tells us that our eyes actually see the Fourier tranform of an object, and it is lens optics that does

30.2.1 Far-Field Approximation of a Linear Array

The vector potential on the xy-plane in the far field, using the sifting property of delta function, yields the following equation for $\mathbf{A}(\mathbf{r})$ using (30.2.2),

$$\mathbf{A}(\mathbf{r}) \cong \hat{z}\frac{\mu Il}{4\pi r}e^{-j\beta r}\iiint d\mathbf{r}'[A_0\delta(x') + A_1\delta(x'-d_1) + \cdots]\delta(y')\delta(z')e^{j\beta\mathbf{r}'\cdot\hat{r}} \tag{30.2.3}$$

In the above, for simplicity, we will assume that the observation point is only on the xy plane, or that $\mathbf{r} = \boldsymbol{\rho} = \hat{x}x + \hat{y}y$ where $\boldsymbol{\rho}$ is the position vector in the xy plane. Thus, $\hat{r} = \hat{x}\cos\phi + \hat{y}\sin\phi$. Also, since the sources are aligned on the x axis, then $\mathbf{r}' = \hat{x}x'$, and $\mathbf{r}'\cdot\hat{r} = x'\cos\phi$. Consequently, $e^{j\beta\mathbf{r}'\cdot\hat{r}} = e^{j\beta x'\cos\phi}$. By so doing, the far field of a linear array is

$$\mathbf{A}(\mathbf{r}) \cong \hat{z}\frac{\mu Il}{4\pi r}e^{-j\beta r}[A_0 + A_1e^{j\beta d_1\cos\phi} + A_2e^{j\beta d_2\cos\phi} + \cdots + A_{N-1}e^{j\beta d_{N-1}\cos\phi}] \tag{30.2.4}$$

Next, for further simplification, we let $d_n = nd$, implying an equally spaced array with distance d between adjacent elements. We next let $A_n = e^{jn\Psi}$, assuming a progressively increasing phase shift between different elements. Such an antenna array is called a *linear phase array*. Thus, (30.2.4) we can identify the array factor $A.F.$ in the above as

$$\begin{aligned} A.F. \cong [1 + e^{j(\beta d\cos\phi+\Psi)} + e^{j2(\beta d\cos\phi+\Psi)} + \cdots \\ + e^{j(N-1)(\beta d\cos\phi+\Psi)}] \end{aligned} \tag{30.2.5}$$

With the simplifying assumptions, the above series can be summed in closed form because it is a series of the form $1 + x + x^2 + x^3 + \cdots + x^{N-1}$. Again, the aggregate pattern of the array is the product of the unit pattern of a lone antenna element multiplied by the array factor.

30.2.2 Radiation Pattern of an Array

The array factor $A.F.$ above(30.2.5) can be summed in closed form using the formula

$$\sum_{n=0}^{N-1} x^n = \frac{1-x^N}{1-x} \tag{30.2.6}$$

Then in the far field,

$$\mathbf{A}(\mathbf{r}) \cong \hat{z}\frac{\mu Il}{4\pi r}e^{-j\beta r}\frac{1-e^{jN(\beta d\cos\phi+\Psi)}}{1-e^{j(\beta d\cos\phi+\Psi)}} \tag{30.2.7}$$

---

the inverse Fourier transform for us. Furthermore, we know that the Fourier transform of an impulse train is an impulse train in our signal processing course. We can extend this to 3D space to ascertain that the Fourier transform of sources located a lattice points become points located at the reciprocal lattice points in the Fourier space [96]. Physicists also call the Fourier space the momentum space because of the association of $\hbar\mathbf{k}$ with momentum.

Ordinarily, as shown previously in (29.1.11), $\mathbf{E} \approx -j\omega(\hat{\theta}A_\theta + \hat{\phi}A_\phi)$. But since $\mathbf{A}$ is $\hat{z}$ directed, we have $A_\phi = 0$. Furthermore, on the xy plane, $E_\theta \approx -j\omega A_\theta = j\omega A_z$. As a consequence,

$$\begin{aligned} |E_\theta| &\cong |E_0| \left| \frac{1 - e^{jN(\beta d \cos\phi + \Psi)}}{1 - e^{j(\beta d \cos\phi + \Psi)}} \right|, \\ &= |E_0| \left| \frac{\sin\left(\frac{N}{2}(\beta d \cos\phi + \Psi)\right)}{\sin\left(\frac{1}{2}(\beta d \cos\phi + \Psi)\right)} \right|, \quad \mathbf{r} \to \infty \end{aligned} \tag{30.2.8}$$

The above can be used to plot the far-field pattern of an antenna array.

Equation (30.2.8) has an array factor that is of the form

$$\frac{|\sin(Nx)|}{|\sin x|}$$

This function, which appears in digital signal processing frequently, is also called the digital sinc function [200]. Again, the reason why this is so is because the far field is proportional to the Fourier transform of the current. The current in this case a finite array of Hertzian dipole, which is a product of a box function and infinite array of Hertzian dipole. The Fourier transform of such a current, as is well known, is the digital sinc.[4]

Plots of $|\sin(3x)|$ and $|\sin x|$ are shown as an example and the resulting $\frac{|\sin(3x)|}{|\sin x|}$ is also shown in Figure 30.4. The function peaks (also called the principal maximum) when both the numerator and the denominator of the digital sinc vanish. Their values can be found by Taylor series expanding both the numerator and the denominator at the point where they both vanish. This happens when $x = n\pi$ for integer n. (Such an apparent singularity is called a removable singularity.)

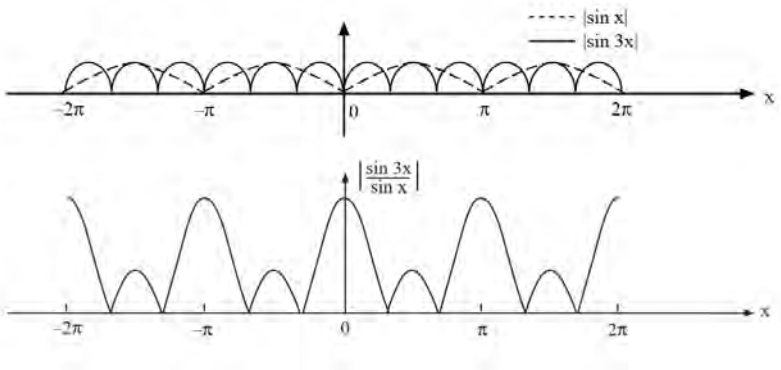


Figure 30.4: Plots of $|\sin x|$, $|\sin 3x|$ of the digital sinc, $\frac{|\sin 3x|}{|\sin x|}$. It peaks at points where both the numerator and denominator vanish.

[4] All good electrical engineers should know that the Fourier transform of an infinite impulse train in the time domain is an impulse train in the frequency domain. Hence, the Fourier transform of an infinitely long array of point sources is an infinitely long array of delta functions in the spectral Fourier space. The Fourier transform of a box function is a sinc function. Hence, the Fourier transform of a finite size array of dipoles is actually the convolution of the sinc function with an infinitely long array of delta function, yielding the digital sinc.

In equation (30.2.8), $x = \frac{1}{2}(\beta d \cos\phi + \Psi)$. We notice that the *maximum* of the array radiation pattern in (30.2.8) would occur if $x = n\pi$, or if

$$\beta d \cos\phi + \Psi = 2n\pi, \qquad n = 0, \pm 1, \pm 2, \pm 3, \cdots \tag{30.2.9}$$

Solving the above gives the location of the maxima of the radiation pattern. The *zeros* or *nulls* of the radiation pattern will occur at $Nx = n\pi$, or

$$\beta d \cos\phi + \Psi = \frac{2n\pi}{N}, \qquad n = \pm 1, \pm 2, \pm 3, \cdots, \quad n \neq mN \tag{30.2.10}$$

For example,

Case I. $\Psi = 0, \beta d = \pi$, principal maximum is at $\phi = \pm\frac{\pi}{2}$. If $N = 5$, nulls are at $\phi = \pm\cos^{-1}\left(\frac{2n}{5}\right)$, or $\phi = \pm 66.4°, \pm 36.9°, \pm 113.6°, \pm 143.1°$. The radiation pattern is seen to form *lobes* of the antenna radiation pattern. The largest lobe is called the main lobe, while the smaller lobes are called side lobes. Since $\Psi = 0$, the radiated fields in the y direction are in phase and the peak of the radiation lobe is in the y direction or in the broadside direction (see Figure 30.5 for the definition of broadside and endfire.). Hence, this is called a *broadside array.* The radiation pattern of such an array is shown in Figure 30.5.

Case II. $\Psi = \pi, \beta d = \pi$, principal maximum is at $\phi = 0, \pi$. If $N = 4$, nulls are at $\phi = \pm\cos^{-1}\left(\frac{n}{2} - 1\right)$, or $\phi = \pm 120°, \pm 90°, \pm 60°$. Since the sources are out of phase by 180°, and $N = 4$ is even, the radiation fields cancel each other in the broadside, but add in the x direction or the end-fire direction. This is called the *endfire array.* Figure 30.6 shows the radiation pattern of such an array.

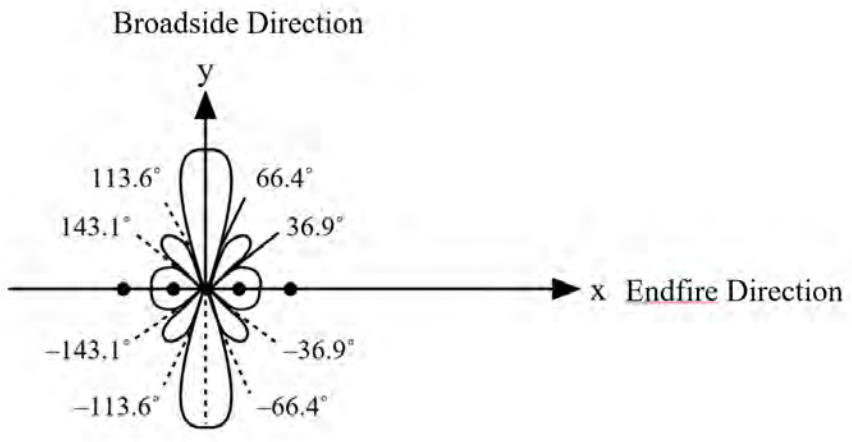


Figure 30.5: The radiation pattern of a five-element broadside array. These terms have descended from war ships. Assuming that the ship is travel in the x direction, then the y direction is the broadside of the ship. The broadside and endfire directions of the array are also labeled.

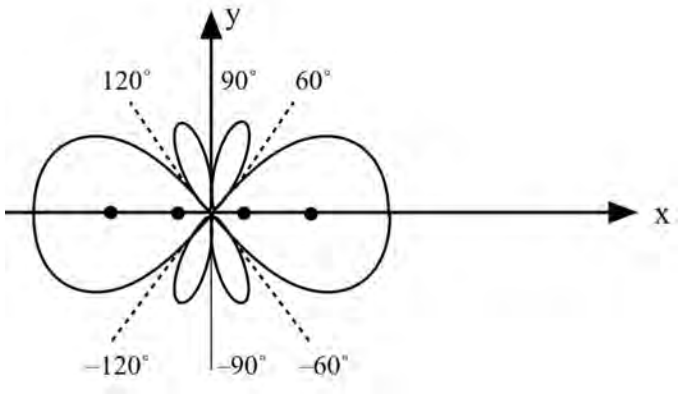


Figure 30.6: By changing the phase of the linear array, the radiation pattern of the antenna array can be changed to become an endfire array as shown.

From the above examples, it is seen that the interference effects between the different antenna elements of a linear array focus the power in a given direction. We can use antenna array to increase the directivity of antennas. Moreover, it is shown that the radiation patterns can be changed by adjusting the spacings of the elements as well as the relative phase shift between them. The idea of antenna array design is to make the main lobe of the pattern to be much higher than the side lobes so that the radiated power of the antenna is directed along the main lobe rather than the side lobes. So side-lobe level suppression is an important goal of designing a highly directive antenna. Also, by changing the phase of the antenna elements in real time, the beam of the antenna can be steered in real time with no moving parts. The antenna array we have studied is of the simplest kind. For instance, the antenna element current amplitudes can be non-uniform, and we will have a non-uniform array.

It is to be noted that the series in (30.2.4) can be thought of as a polynomial if the array is equally spaced, and the phases for A_n properly chosen. A typical term in (30.2.4) is of the form

$$A_n e^{j\beta d_n \cos\phi} = A_n e^{j\beta nd\cos\phi} = A_n \left(e^{j\beta d\cos\phi}\right)^n \tag{30.2.11}$$

or $A_n x^n$.

Sometimes, this polynomial can be summed in closed form. These are the cases in the binomial arrays and Chebyshev arrays, and their radiation patterns can be easily found, and amenable to engineering designs.

30.3 Validity of the Far-Field Approximation

In making the far-field approximation in (30.2.3), it will be interesting to ponder when the far-field approximation is valid? That is, when we can approximate

$$e^{-j\beta|\mathbf{r}-\mathbf{r}'|} \approx e^{-j\beta r + j\beta \mathbf{r}'\cdot\hat{r}} \tag{30.3.1}$$

to arrive at (30.2.2) and (30.2.3). This is especially important because when we integrate over $\mathbf{r}'$, it can range over large values especially for a large array. In this case, $\mathbf{r}'$ can be as large as $(N-1)d$. The above approximation is important also because it tells when the field generated by an array antenna becomes a spherical wave.

To answer this question, we need to study the approximation in (30.3.1) more carefully. First, we have

$$|\mathbf{r}-\mathbf{r}'|^2 = (\mathbf{r}-\mathbf{r}')\cdot(\mathbf{r}-\mathbf{r}') = r^2 - 2\mathbf{r}\cdot\mathbf{r}' + r'^2 \tag{30.3.2}$$

We can take the square root of the above to get

$$|\mathbf{r}-\mathbf{r}'| = r\left(1 - \frac{2\mathbf{r}\cdot\mathbf{r}'}{r^2} + \frac{r'^2}{r^2}\right)^{1/2} \tag{30.3.3}$$

The above is exact so far with no approximation. Next, we use the Taylor series expansion to get, for small x, that

$$(1+x)^n \approx 1 + nx + \frac{n(n-1)}{2!}x^2 + \cdots \tag{30.3.4}$$

or that

$$(1+x)^{1/2} \approx 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \cdots \tag{30.3.5}$$

We can apply this approximation by letting

$$x \cong -\frac{2\mathbf{r}\cdot\mathbf{r}'}{r^2} + \frac{r'^2}{r^2} \tag{30.3.6}$$

in (30.3.3). To this end, we arrive at[5]

$$|\mathbf{r}-\mathbf{r}'| \approx r\left[1 - \frac{\mathbf{r}\cdot\mathbf{r}'}{r^2} + \frac{1}{2}\frac{r'^2}{r^2} - \frac{1}{2}\left(\frac{\mathbf{r}\cdot\mathbf{r}'}{r^2}\right)^2 + \cdots\right] \tag{30.3.7}$$

In the above, we have not kept every term of the x^2 terms by assuming that $r'^2 \ll \mathbf{r}'\cdot\mathbf{r}$, and terms much smaller than the last term in (30.3.7) can be neglected.

We can multiply out the right-hand side of the above to further arrive at

$$\begin{aligned} |\mathbf{r}-\mathbf{r}'| &\approx r - \frac{\mathbf{r}\cdot\mathbf{r}'}{r} + \frac{1}{2}\frac{r'^2}{r} - \frac{1}{2}\frac{(\mathbf{r}\cdot\mathbf{r}')^2}{r^3} + \cdots \\ &= r - \hat{r}\cdot\mathbf{r}' + \frac{1}{2}\frac{r'^2}{r} - \frac{1}{2r}(\hat{r}\cdot\mathbf{r}')^2 + \cdots \end{aligned} \tag{30.3.8}$$

The last two terms in the last line of (30.3.8) are of the same order.[6] Moreover, their sum is bounded by $r'^2/(2r)$ since $\hat{r}\cdot\mathbf{r}'$ is always less than r'. Hence, using the above in (30.3.1), the far

[5]The art of making such approximation is called perturbation expansion [48].

[6]The math parlance for saying that these two terms are approximately of the same magnitude as each other.

field approximation is valid if

$$\beta \frac{r'^2}{2r} \ll 1 \tag{30.3.9}$$

In the above, β is involved because the approximation has to be valid in the exponent of (30.3.1), namely $\exp(-j\beta|\mathbf{r}-\mathbf{r}'|)$ where β multiplies $|\mathbf{r}-\mathbf{r}'|$ or its approximation. If (30.3.9) is valid, then

$$e^{j\beta \frac{r'^2}{2r}} \approx 1$$

and thus, the first two terms on the right-hand side of (30.3.8) suffice to approximate $|\mathbf{r}-\mathbf{r}'|$ on the left-hand side, which are the two terms we have kept in the far-field approximation.

30.3.1 Rayleigh Distance

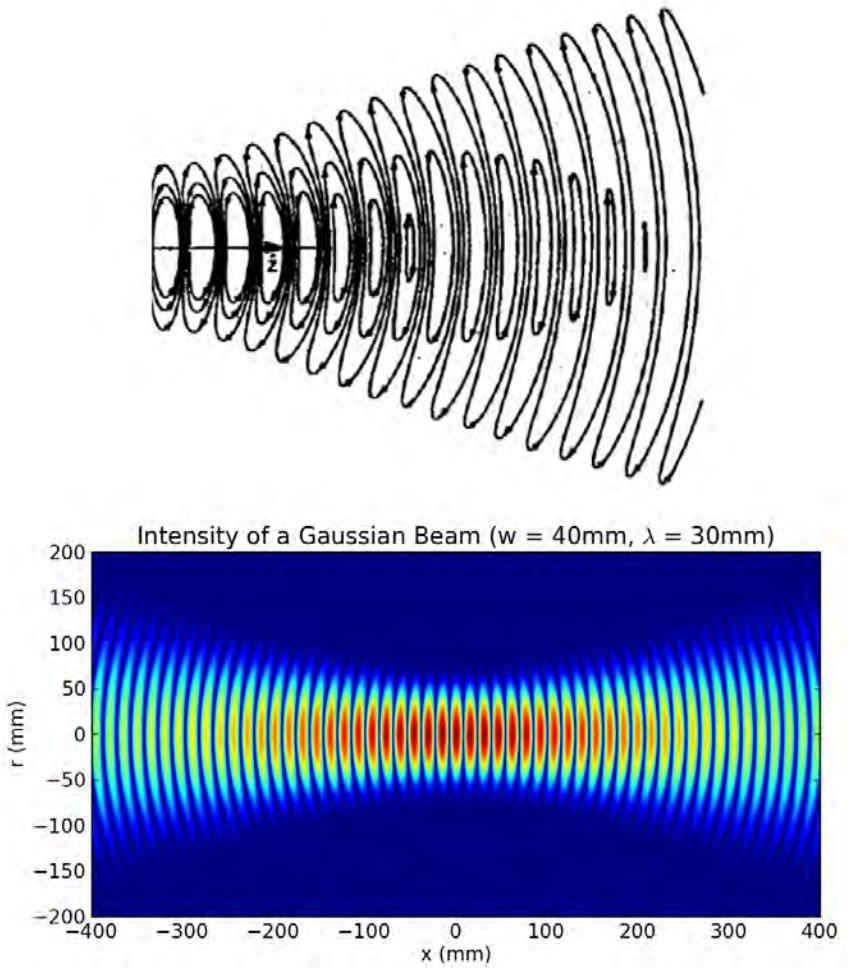


Figure 30.7: In the top figure, which is the right half of a Gaussian beam, it displays the physics of the near field, the Fresnel zone, and the far zone as one moves from the left to the right in the picture. In the far zone, the field behaves like a spherical wave (courtesy of [89], the bottom figure is courtesy of I. Okhmatovskii).

If we have an infinite time-harmonic current sheet in the xy plane, it can be shown that by matching boundary conditions, it will launch plane waves on both sides of the current sheet propagating in the z direction [34][p. 652].

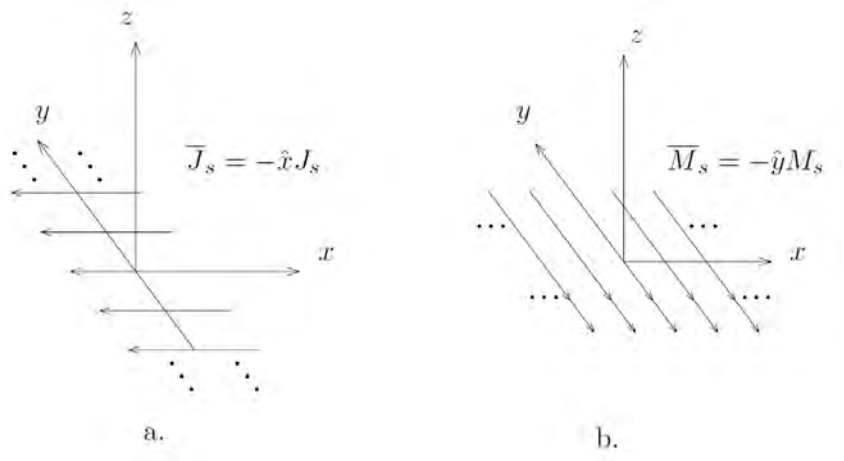


Figure 30.8: An infinite electric or magnetic current sheet, by the equivalence principle, can launch a perfect plane wave. The plane waves are propagating above and below the current sheet. [34, p. 653]).

Thus if we have an aperture antenna like the opening of a waveguide that is much larger than the wavelength, it will launch a wave that is almost like a plane wave from it. We can use Huygens' original hypothesis that the current wavefront can be predicted by putting point sources at the previous wavefront. This is also vindicated by our derivation of Huygens' principle in the previous lecture. Thus, when a wave field leaves a large aperture antenna, it can be approximately described by a plane wave, and later, by a Gaussian beam [89] (see Figure 30.7).

As seen in the above figure, near to the antenna aperture, or in the near zone, it is approximately a plane wave with wave fronts parallel to the aperture surface. Far from the antenna aperture, or in the far zone, the field behaves like a spherical wave, with its typical wave front. In between, we are in the Fresnel zone. A Gaussian beam describes the morphing of a finite size plane wavefront to a spherical wavefront.

Consequently, after using that $\beta = 2\pi/\lambda$, for the far-field approximation to be valid, we need (30.3.9) to be valid, or that

$$r \gg \frac{\pi}{\lambda} r'^2 \tag{30.3.10}$$

If the aperture of the antenna is of radius W, then $r' < r_{\max} \cong W$ and the far field approximation is valid if

$$r \gg \frac{\pi}{\lambda} W^2 = r_R \tag{30.3.11}$$

If r is larger than this distance, then the far-field approximation is valid, and an antenna beam behaves like a spherical wave: it starts to diverge. This distance r_R is also known as the Rayleigh distance. In other words, after this distance, the wave from a finite size source resembles a spherical

wave which is diverging in all directions (see Figure 30.7). Also, notice that the shorter the wavelength λ, the larger is this distance before the far field approximation can be made.

This also explains why a laser pointer works. A laser pointer light can be thought of radiation from a finite size source located at the aperture of the laser pointer as shall be shown using equivalence theorem later. The laser pointer beam remains collimated for quite a distance, before it becomes a divergent beam or a beam with a spherical wave front.

In some textbooks [34], it is common to define acceptable phase error to be $\pi/8$. The Rayleigh distance is the distance beyond which the phase error is below this value. When the phase error of $\pi/8$ is put on the right-hand side of (30.3.9), one gets

$$\beta \frac{r'^2}{2r} \approx \frac{\pi}{8} \tag{30.3.12}$$

Using the approximation, the Rayleigh distance is defined to be

$$r_R = \frac{2D^2}{\lambda} \tag{30.3.13}$$

where $D = 2W$ is the diameter of the antenna aperture. Learning the demarcation distance between near zone, Fresnel zone, and the far zone is a concept is important both optics and microwave.

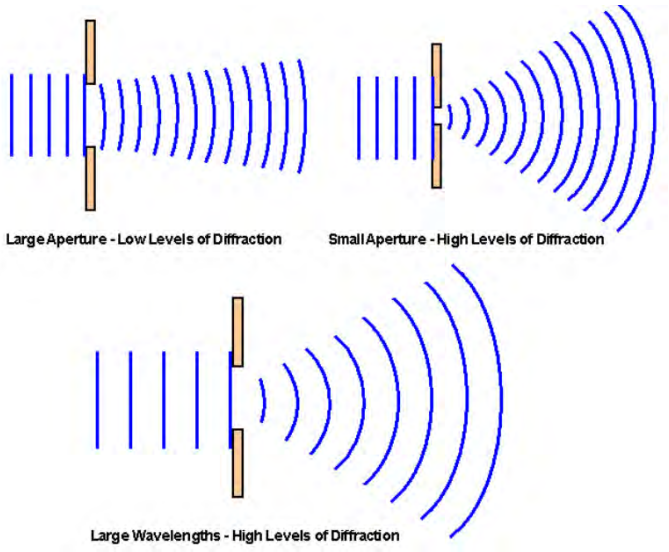


Figure 30.9: Diffraction of a plane wave by apertures of different sizes. The laser pointer works like diffraction of short wavelength plane waves by a large aperture. The beam remains like a plane wave after diffraction, and hence it is collimated for a Rayleigh distance before beam divergence set in (courtesy of imagen-estilo.com).

30.3.2 Near Zone, Fresnel Zone, and Far Zone

Therefore, when a source radiates, the radiation field is divided into the near zone, the Fresnel zone, and the far zone (also known as the radiation zone, or the Fraunhofer zone in optics). The Rayleigh distance is the demarcation boundary between the Fresnel zone and the far zone. The larger the aperture of an antenna array is, the further one has to be in order to reach the far zone of an antenna from (30.3.13). This distance becomes larger too when the wavelength is short. In the far zone, the far field behaves like a spherical wave, and its radiation pattern is proportional to the Fourier transform of the current.

In some sources, like the Hertzian dipole, in the near zone, much reactive energy is stored in the electric field or the magnetic field near to the source. This near zone receives reactive power from the source, which corresponds to instantaneous power that flows from the source, but is return to the source after one time harmonic cycle. Thus, reactive power corresponds to energy that sloshes back and forth between the source and the near field. Hence, a Hertzian dipole has input impedance that looks like that of a capacitor, because much of the near field of this dipole is the electric field.

The field in the far zone carries power that radiates to infinity. But the near field stores energy and carries reactive power that does not propagate to infinity. As a result, the field in the near zone decays rapidly. Furthermore, it sustains reactive power that needs to be exchanged with the source. On the other hand, the field in the far zone decays as $1/r$ for energy conservation. Moreover, the far field is no longer bound to the source, as it convects energy to infinity.

Exercises for Lecture 30

Problem 30-1:

(i) Go through the lecture notes and derive the expression (30.2.8).

(ii) Repeat Case II of the same lecture notes, but with $N = 5$, and plot the far field pattern of this new array.

(iii) Find the leading order approximation, up to the quadratic term, of the expression $|\mathbf{r} - \mathbf{r}'|$ when $r' \ll r$. In other words, rederive equation (30.3.8) and reconfirm the definition of Rayleigh distance in (30.3.11). Also, derive the Rayleigh distance defined in (30.3.13).

(iv) Explain why a laser pointer beam remains collimated after the light beam has left the aperture of the laser pointer.

Chapter 31

Different Types of Antennas—Heuristics

We have studied different closed form solutions and approximate solutions to Maxwell's equations. Examples of closed form solutions are found in transmission lines, waveguides, resonators, and dipoles. Examples of approximate solutions are found in circuit theory and far field approximations. These solutions offer us insights into the physical behaviour of electromagnetic fields, and also the physical mechanisms as to how things work. These physical insights often inspire us to develop new designs.

Fortunately for us, Maxwell's equations are valid from sub-atomic lengthscales to galactic lengthscales. In vacuum, they have been validated to extremely high accuracy (see Section 1.1). Furthermore, in the last few decades since the 1960s, very many numerical solutions provided by commercial software have been possible for solving Maxwell's equations of complex structures. This field of solving Maxwell's equations numerically is known as computational electromagnetics (CEM) which shall be discussed later in this course. Many commercial software are now available to solve Maxwell's equations to high fidelity. Therefore, design engineers these days do not require advanced knowledge of math and physics, and the solutions of Maxwell's equations can be obtained by learning how to use these commercial software. This is a boon to many design engineers: by running these software with cut-and-try engineering, wonderful systems can be designed. The art of electromagnetic design using simulation before the actual hardware is made is known as virtual prototyping. They have replaced expensive cut-and-try experiments.

It used to be said that if we lock 100 monkeys in a room and let them punch at the 100 keyboards, they will never type out Macbeth nor Hamlet. But with 100 engineers trained with good physical insight, when locked up in a room with commercial software, with enough time and patience, they can come up with wonderful designs of different electromagnetic systems. Antenna design now is mainly driven by heuristics and cut-and-try engineering for virtual protyping. Therefore, we will discuss the functions of different antennas heuristically in this lecture.

31.1 Resonance Tunneling in Antenna

We realize the power of resonance enhancement when we were young by playing on a swing in the park. By pumping the swing at its resonance frequency, we can cause it to swing at a large amplitude without a Herculean effort. A simple antenna like a short dipole behaves like a Hertzian dipole with an effective length. A short dipole has an input impedance resembling that of a capacitor. Hence, it is difficult to drive current into the antenna unless other elements are added. Hertz was clever by using two metallic spheres attached to the stem of the dipole to increase the current flow. A large current flow on the stem of the antenna makes the stem resemble an inductor. Thus, the end-cap capacitances and the stem inductance together act like a resonator enhancing the current flow on the antenna.

Some antennas are deliberately built to resonate with its structure to enhance its radiation. A half-wave dipole is such an antenna as shown in Figure 31.1 [194]. These antennas use resonance tunneling to increase the currents on them, and thus to enhance their radiation efficiencies. A half-wave dipole can also be thought of as a flared open transmission line in order to make it radiate. The transmission line can be gradually morphed from a quarter-wavelength transmission line as shown in Figure 31.1. A transmission line is a poor radiator, because the electromagnetic energy is trapped between two pieces of metal. But a flared transmission line can radiate its field to free space more easily. (The dipole antenna, though a simple device, has been extensively studied by King [201].[1])

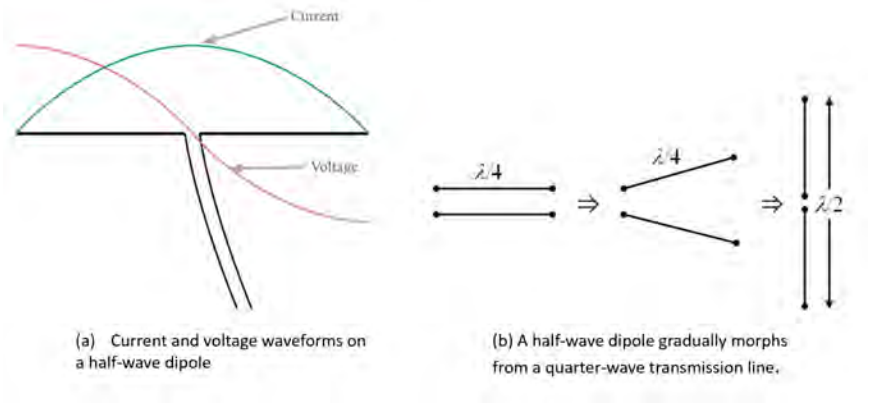


Figure 31.1: A half-wave dipole can be thought of as a resonator with radiation loss. It can be thought of as a quarter-wavelength transmission line that is gradually opened up or flared (courtesy of electronics-notes.com).

One can also think of a long piece of a wire as a waveguide. It is called a *Goubau line* as shown in Figure 31.2, which can be thought of the limiting case of a coaxial cable where the outer conductor is gradually moved to be infinitely far away [134]. The wave is weakly guided since it

[1] He has reputed to have graduated more than 100 PhD students studying the dipole antenna.

now can shed energy to infinity. The behavior of a wire as a Goubau line waveguide can be used to explain heuristically why a half-wave dipole resonates when it is about half wavelength.

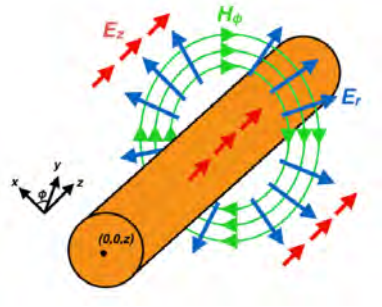


Figure 31.2: The electromagnetic field around a Goubau line (courtesy of [134]). The field resembles that of a coaxial line with the outer conductor (or ground) gradually moved to infinity. It can be thought of as a quasi-transmission line. When applied to a half-wave dipole, the mode is quasi-guided on the wire, and is torn off from the wire antenna it reaches its ends. Then it is launched into space at the end of the line.

Figure 31.3: A Yagi-Uda antenna was invented by heuristics and great physical insight in 1926 even before we had computers. The principal element of the antenna is the folded dipole with four times the radiation resistance of a half-wave dipole. When an array of wire dipole antenna, each of which is less than half a wavelength, the array acts as a waveguide, or a director. When the wire antenna is slightly more than half a wavelength, the wire dipole ceases to be a waveguide, and it acts as a reflector [202]. Therefore, the antenna radiates predominantly in one direction (courtesy of Wikipedia [203]).

A folded dipole is often used to alter the input impedance of a dipole antenna [204]. Even though it can have a resonant frequency lower than that of a normal dipole, the lowest resonant mode does not radiate well due to destructive cancellation. The mode that radiates well has the

same resonant length as an unfolded dipole. It has a radiation resistance about four times that of a half-wave dipole of similar length which is about 300 ohms. This is equal to the characteristic impedance of a twin-lead transmission line [205]. Figure 31.3 shows a Yagi-Uda antenna driven by a folded dipole. This antenna was very popular and adorned the roof of every household before high frequency cable modems brought broadband signals to our homes, which happened around mid 1990s.

A Yagi-Uda antenna is also another interesting invention. It was invented in 1926 by Yagi and Uda in Japan by plainly using physical intuition [202]. Physical intuition was a tool of engineers of yesteryears while modern engineers prefer to use sophisticated computer-aided design (CAD) software. Nevertheless, physical intuition is still important. The principal driver element of the antenna is the folded dipole. Surprisingly, the array of dipole elements, whose length is slightly less than a half wavelength, in front of the driver element are acting like a waveguide in space, while the sole element at the back, slightly larger than a half wavelength, acts like a reflector. Therefore, the field radiated by the driver element will be directed toward the front of the antenna. Thus, this antenna has higher directivity than just a stand alone dipole. Due to its simplicity, this antenna has been made into nano-antennas which operate at optical frequencies [206].

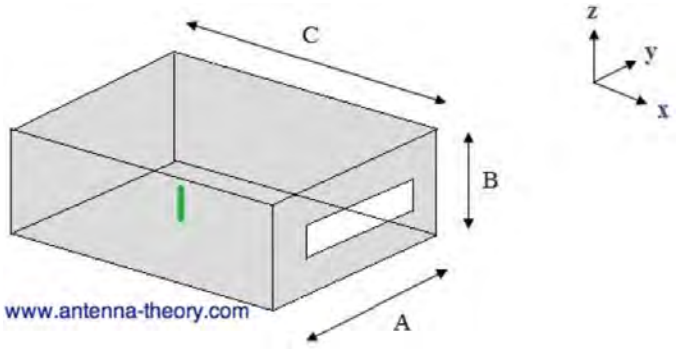


Figure 31.4: A cavity-backed slot antenna radiates well using the physics of resonant tunneling. When the small dipole radiates close to the resonant frequency of the cavity, the field strength is strongly enhanced inside the cavity, and hence around the slot. This makes the slot into a good radiator (courtesy of antenna-theory.com).

Slot antenna is a simple antenna to make [207]. To improve the radiation efficiency of slot antenna, it is made to radiate via a cavity. A cavity-backed slot antenna that uses such a concept is shown in Figure 31.4. A small dipole with poor radiation efficiency is placed inside the cavity. When the operating frequency is close to the resonant frequency of the cavity, the field strength inside the cavity becomes very strong, and much of the energy can leak from the cavity via the slot on its side. This makes the antenna radiate more efficiently into free space compared to just the small dipole alone, due to the physics of resonant tunneling.

Another antenna that resembles a cavity backed slot antenna is the microstrip patch antenna (or just called a patch antenna). This is shown in in Figure 31.5. This antenna also radiates efficiently by resonant tunneling. Roughly, when L (see left of Figure 31.5) is half a wavelength,

the patch antenna resonates. It is similar to the resonant frequency of a transmission line with open circuit at both ends. The current sloshes back and forth across the length of the patch antenna along the L direction. The fringing electric fields at the two ends can be thought of as equivalent magnetic currents that radiate in phase in the direction normal to the patch.

The second design (right of Figure 31.5) has an inset feed. It allows the antenna to resonate at a lower frequency because the current has to go through a tortuos path, and has a longer path to slosh through when it is at resonance. Then it resonates at a lower frequency meaning that the antenna can be made smaller.

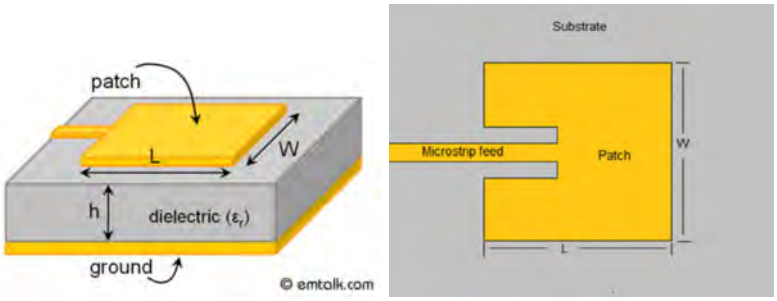


Figure 31.5: A microstrip patch antenna also radiates well when it resonates. The patch antenna resembles a cavity resonator with magnetic wall [208]. Again, it uses the physics of resonant tunneling to enhance its radiation efficiency. The notch increases the length of the current flow, and reduce the resonant frequency of the antenna (courtesy of emtalk.com).

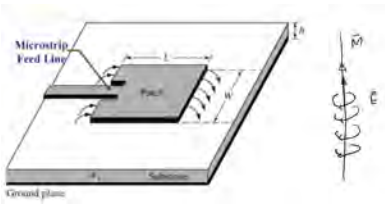


Figure 31.6: A microstrip patch antenna showing fringing electric fields at the two apertures. The fringing electric fields can be thought of as equivalent magnetic currents. The magnetic currents have the same E fields, and they can be used to model the radiation of the patch antenna. (courtesy of everthingRF.com).

31.2 Horn Antennas

The impedance of free space is 377 ohms while that of most transmission line is about 50 ohms. This impedance mismatch can be mitigated by using a flared horn (see Figure 31.7) [209]. The

gradual transition region allows the wave to travel from a region of low impedance to a region of high impedance with reduced reflection.

One can think that the characteristic impedance is $Z_0 = \sqrt{L/C}$ of a transmission line if it is made of two pieces of metal. As the horn flares, C becomes smaller, increasing its characteristic impedance to get close to that of free space which is about 377 siemens. This allows for better impedance matching from the source to free space. It is similar to the quarter wave transformer for matching the characteristic impedance Z_0 of a line to a load with impedance Z_L. The requirement is that the quarter wave transformer has an impedance given by $Z_T = \sqrt{Z_0 Z_L}$, which is the geometrical mean of the two impedances.

A corrugated horn, as we have discussed previously in a circular waveguide in Section 24.1.1, discourages current flows in the non-axial symmetric modes. The reason is that the axial symmetric modes have only circumferential currents while the non-axial symmetric modes have axial currents. The corrugation impedes the flow of axial currents, and hence, discourages the propagation of the non-axial symmetric modes. On the contrary, it encourages the propagation of the axial symmetric TE_{01} mode in the circular waveguide or the circular horn antenna. Because this mode is axially symmetric, this antenna can radiate fields that are axially symmetric giving rise to an axially symmetric radiation pattern [210, 211, 212].

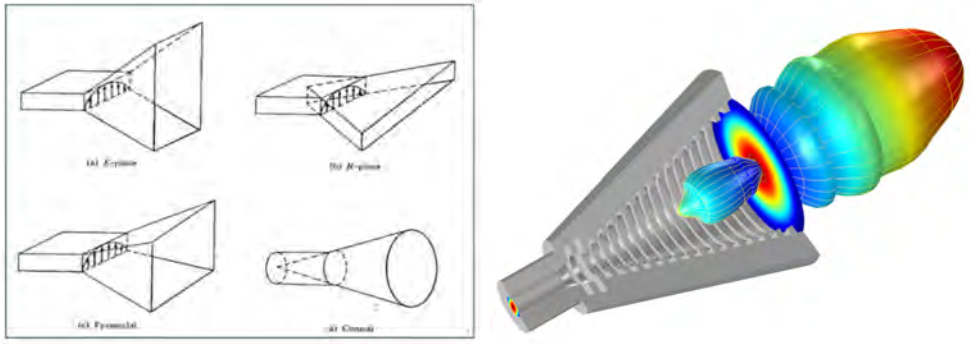


Figure 31.7: A horn antenna works with the same principle as the biconical antenna. Its flared horn changes the waveguide impedance so as to match the impedance of a waveguide to the impedance of free space. The lower figure is that of a corrugated circular horn antenna. The corrugation enhances the propagation of the TE_{01} mode in the circular waveguide, and thus it enhances the cylindrical symmetry of the mode and the radiation field (courtesy of tutorialpoints.com and comsol.com).

A Vivaldi antenna (invented by P. Gibson in 1978 [213]), is shown in Figure 31.8.[2] It is also called a notched antenna, and the two pieces of metal act as a coplanar waveguide (see [157][p. 4]) It works by the same principle to gradually match the impedance of the source to that of free space. But such a gradually flared coplanar waveguide (flared horn) has the element of a frequency independent antenna. The low frequency component of the signal will radiate from the wide end of the flared notch, while the high frequency component will radiate from the narrow end of the

[2]He must have loved the musician Vivaldi so much:)

notch. Thus, this antenna can radiate effectively over a broad range of frequencies, giving the antenna a broad bandwidth performance. It is good for transmitting a pulsed signal which has a broad frequency spectrum.

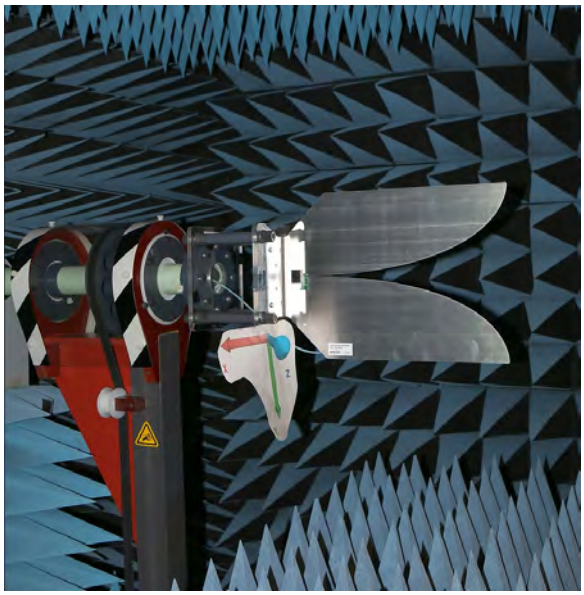


Figure 31.8: A Vivaldi antenna, also called a notched antenna, works like a horn antenna, but uses very little metal and is lightweight. Hence, it is cheap to build, and its flared notch makes it broadband. It is broadband because the high frequency signals can radiate from the narrow part of the notch, and the lower frequency signals can radiate from the wider part of the notch (courtesy of Wikipedia [214, 215]).

31.3 Quasi-Optical Antennas

High-frequency or short-wavelength electromagnetic field behaves like light ray as in optics. Therefore, many high-frequency antennas are designed based on the principle of ray optics. A reflector antenna is such an antenna as shown in Figure 31.9. The reflector antenna in this case is a Cassegrain design [216][3] where a sub-reflector is present. This allows the antenna to be fed from behind the parabolic dish where the electronics can be stored and isolated as well. Reflector antennas [218] are prevalent in radio astronomy and space exploration due to their high directivity and sensitivity needed for low signals. Moreover, due to their large size compared to wavelength, they have a large effective aperture or area.

[3]The name came from an optical telescope of similar design [217]

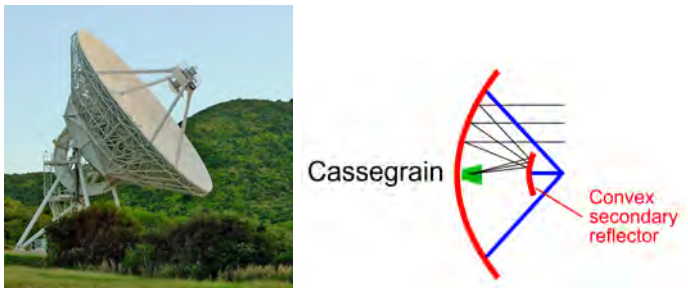


Figure 31.9: The left picture of an NRAO radio telescope antenna of Cassegrain design in Virginia, USA (courtesy of Britannica.com). The right is the detail of the Cassegrain design (courtesy of rev.com). Its design is inspired by ray optics. In fact, the Cassegrain design was first used in optical telescope.

Another recent invention is the reflectarray antenna [219, 220] which is very popular. One of them is shown in Figure 31.10. Due to recent advent in simulation technology, complicated structures can be simulated on a computer, including one with a complicated surface design. Patch elements can be etched onto a flat surface as shown, giving it an effective impedance that is spatially varying, making it reflect like a curved surface. Such a surface is known as a metasurface [221, 222]. Its flat structure can greatly economize on the space usage compared to a reflector antenna.

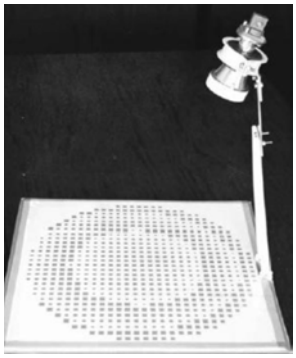


Figure 31.10: A reflectarray where the reflector is a flat surface. Patches are unequally spaced to give the array the focussing effect. No closed form solution exists for such a reflectarray, but due to advancement in computational electromagnetics (CEM), it can be simulated or virtual-prototyped in a computer (courtesy of antenna-theory.com).

Another quasi-optical antenna is the lens antenna as shown in Figure 31.11 [223]. The design of this antenna follows lens optics, and is only valid when the wavelength is very short compared to the curvature of the surfaces. In this case, reflection and transmission at a curved surface is similar to that of a flat surface. This is called the tangent-plane approximation of a curved surface,

and is valid at high frequencies.

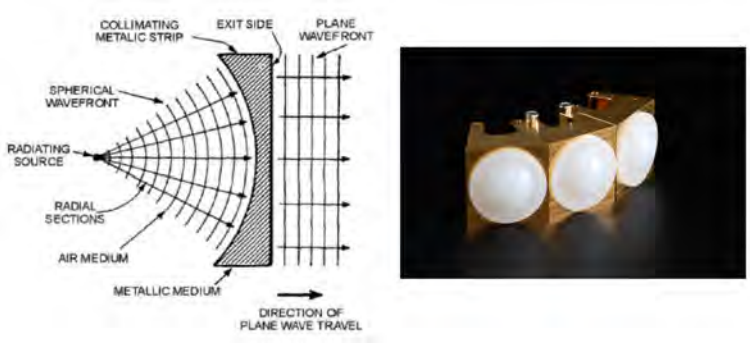


Figure 31.11: The left figure shows a lens antenna where the lens is made of artificial dielectrics made from metallic strips (courtesy of electriciantutoring.tpub.com). The right figure shows some dielectric lens at the aperture of an open waveguide to focus the microwave exiting from the waveguide opening (courtesy of micro-radar.de). When the wavelength is much smaller than the size of the structure, a wave behaves like a ray: quasi-optical concepts can be used to design the antenna.

31.4 Small Antennas

Small antennas are in vogue these days due to the advent of the cell phone, and the importance of economizing on the antenna size due to miniaturization requirements. Also, the antennas should have enough bandwidth to accommodate the signals from different cell phone companies, which use different carrier frequencies. An interesting small antenna is the PIFA (planar inverted F antenna because it is shaped like an F) shown in Figure 31.12 [224]. Because it is shorted at one end and open circuit at the other end, it acts like a quarter wavelength resonator, making it substantially smaller. But the resonator has a low Q because of the "slots" or "openings" around it from whom energy can leak giving rise to radiation loss. The low Q gives this antenna a broader bandwidth. Because it is shorted at one end, its driving point impedance can be changed by altering the feed location.

An interesting small antenna is the U-slot antenna shown in Figure 31.13 [225, 226]. Because the current is forced to follow a longer tortuous path by the U-slot, it can resonant with a longer wavelength (lower frequency) and hence, can be made smaller compared to wavelength. In order to give the antenna a larger bandwidth, its Q is made smaller by etching it on a thick dielectric substrate (shown as the dielectric material region in the figure). But feeding it with a longer probe will make the bandwidth of the antenna smaller, due to the larger inductance of the probe.[4] An ingenious invention is to use an L probe [227]. The L probe has an inductive part as well as a capacitive part. Their reactance cancel each other, allowing the electromagnetic energy to tunnel

[4]Remember that larger inductance implies more store magnetic field energy, and hence, the higher Q of the system.

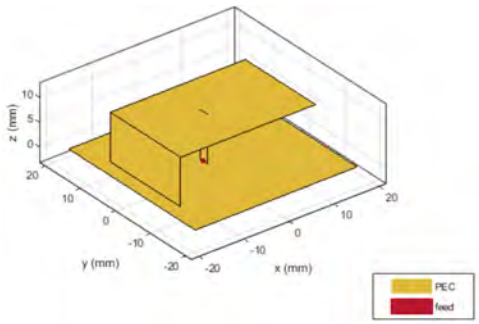


Figure 31.12: A PIFA (planar inverted F antenna) is compact, broadband, and easy to fabricate. It behaves like a quarter wavelength transmission line resonator. It is good for cell phone antennas due to its small size (courtesy of Mathworks).

through the antenna, making it a better radiator.

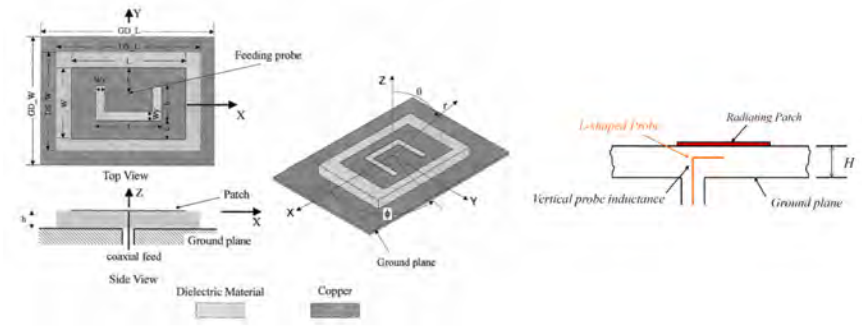


Figure 31.13: The left figure shows a U slot patch antenna design. The right figure shows a patch antenna fed by an L probe with significant increase in bandwidth. The capacitive coupling of the L probe together with its inductance could form an LC tank circuit to help with the impedance matching of the antenna (courtesy of K.M. Luk) [227].

Another area where small antennas are needed is in RFID (radio frequency identification) tag [228]. Since tags are placed outside the packages of products, e.g., in a warehouse, an RFID tag has a transmit-receive antenna that can communicate with the external world.[5] The communication is done through an RFID reader. The RFID reader can talk to a small computer chip embedded in the tag where data about the package can be stored. Thus, an RFID reader can quickly and remotely communicate with the RFID tag to retrieve information about the package. Such a small antenna design for RFID tag is shown in Figure 31.14. It uses image theorem (that we shall learn later) so that the antenna can be made half as small. Then slots are cut into the radiating patch, so

[5] A lower frequency version of it is used in credit and ID cards.

that the current follows a longer path. This lowers the resonant frequency of the antenna, allowing it to be made smaller. The take-home message here is that to make an antenna a few times smaller than a wavelength to resonate, the current on the antenna has to flow through a tortuous path. In this manner, the antenna can be made a few times smaller than the wavelength.

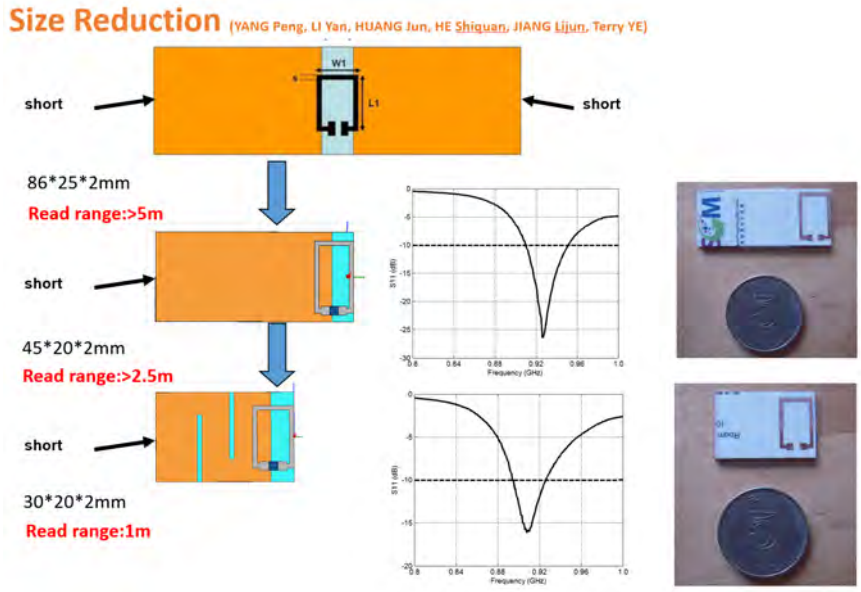


Figure 31.14: Some RFID antennas designed at The University of Hong Kong in collaboration with LSCM (Logistics and Supply Chain Multitech R&D Center, Hong Kong). The image theorem together with the use of tortuous current flow pattern are used to miniaturize the antennas (courtesy of P. Yang, Y. Li, J. Huang, L.J. Jiang, S.Q. He, T. Ye, and W.C. Chew).

An RFID reader can be designed to read the information from a batch of vials (or test tubes) containing different chemicals or medicine. Reading the info from a large batch of vials is important in the pharmaceutical industry. Hence, a large loop antenna is needed but at a sufficiently high frequency (for large bandwidth). However, a loop antenna, if we look at a piece of wire as a Goubau transmission line [134], will have resonant frequencies. When a loop antenna resonates, the current is non-uniform on it just as in a transmission line. This happens at higher frequencies. (Fundamentally, this comes from the retardation effect of electromagnetic field). It will result in a non-uniform magnetic field inside the loop defeating the design of the RFID reader.

One way to view the physics of non-uniform current is that a line of wire can be viewed as a quasi-transmission line. It can be modeled with the lumped element model as shown in Figure 15.3. The series inductance in the lumped element model can be series-loaded with capacitance to reduce its series impedance, to increase the phase velocity of a wave on such a line. This will help the current around a loop to be equi-phase and hence, more uniform [229]. In this way, the

voltage and phase are equalized between points on the loop and become uniform across the loop so is the current. Therefore, one way to enable a uniform current in a large loop is to capacitively load the loop. This will ensure a constant phase, or a more uniform current around the loop, and hence, a more efficient reader. Such a design is shown in Figure 31.15.

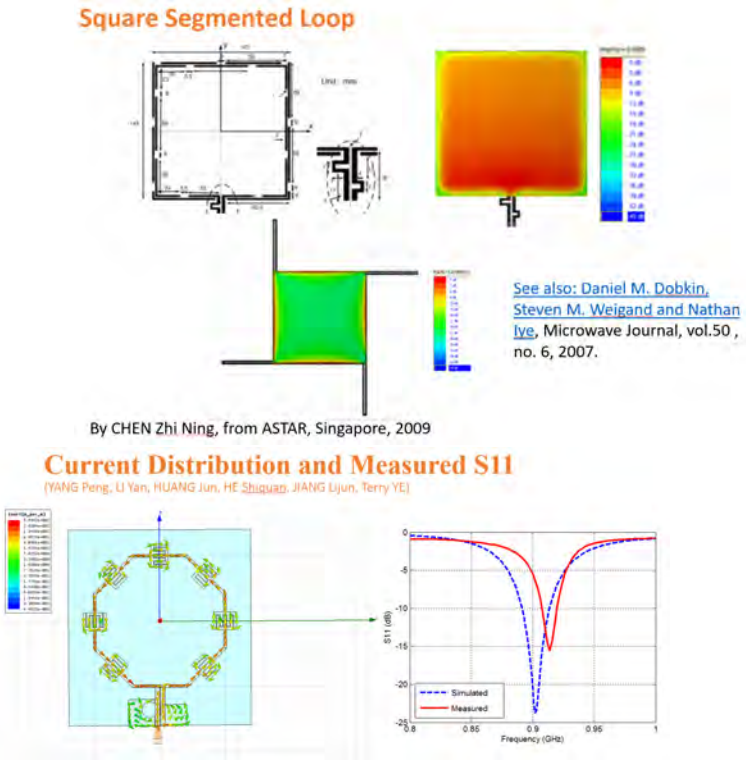


Figure 31.15: The top figure shows a RFID reader designed by [230] using capacitively loaded inductor loop. The bottom figure shows simulation and measurement done at The University of Hong Kong and LSCM (courtesy of Z.N. Chen [230] and P. Yang, Y. Li, J. Huang, L.J. Jiang, S.Q. He, T. Ye, and W.C. Chew). The series impedance is made to vanish to reduce the voltage drop across it, and hence, reduces the phase shift.

Exercises for Lecture 31

Problem 31-1:

(i) Explain why a folded dipole has a radiation resistance four times that of an unfolded dipole **(Hint: the answer can be found in wiki. But please regurgitate it, understand the logic before you provide an answer.)**

(ii) Explain how a cavity backed slot antenna works.

(iii) Explain why the corrugated horn antenna produces an axially symmetric radiation pattern.

(iv) Explain how a lens antenna works.

(v) Explain why PIFA is smaller than a half-wave dipole.

(vi) Explain how you would make the current uniform on a large loop antenna. Use the lumped element model, namely, and show that the phase velocity can go to infinity. You can use (15.2.11) to find the phase velocity of a capacitively loaded transmission line, and explain how you would choose the capacitive loading so that the phase velocity is infinite. You can assume the lossless case so that R and G in (15.2.11) are zero.

Chapter 32

Shielding, Image Theory

The physical mechanism of electromagnetic shielding and electromagnetic image theory (also called image theorem) go hand in hand. They work by the moving of charges around so as to cancel the impinging fields. By understanding simple cases of shielding and image theory, we can gain enough insight to solve some real-world problems. For instance, the art of shielding is very important in the field of electromagnetic compatibility (EMC) and electromagnetic interference (EMI). In the modern age where we have more electronic components working side by side, in a very compact environment, e.g. inside a cell phone (see Figure 32.1), EMC/EMI become an increasingly challenging issue. Due to the complexity of these problems, they have to be solved using heuristics with a high dosage of physical insight and experience.

Figure 32.1: The compactness of the cell phone components make urgent use of EMC/EMI (electromagnetic compatibility/electromagnetic interferece) knowledge instrumental in the design of cell phones. Clever use of shielding is necessary to prevent interferences between different components. Compatibility means that even when each component works well in isolation, they continue to work well when brought together.

32.1 Shielding

We can understand shielding by understanding how electric charges move around in a conductive medium. They move around to shield out the electric field, or cancel the impinging field inside the conductor. There are two cases to consider: the static case and the dynamic case. The physical arguments needed to understand these two cases are very different. Moreover, since there are no magnetic charges around, the shielding of magnetic field is very different from the shielding of electric field, as shall be seen below.

32.1.1 A Note on Electrostatic Shielding

We begin with the simple case of electrostatic shielding. For electrostatic problems, a conductive medium suffices to produce surface charges that shield out the electric field from the conductive medium. If the electric field is not zero, then since $\mathbf{J} = \sigma\mathbf{E}$, the electric current inside the conductor will keep flowing. The current will produce charges on the surface of the conductor to cancel the impinging field, until inside the conductive medium, $\mathbf{E} = 0$. In this case, electric current ceases to flow in the conductor. This is the quiescent limit.

In other words, when the field reaches the quiescent state, the charges have to redistribute themselves so as to shield out the electric field, and that the total internal electric field, $\mathbf{E} = 0$ at equilibrium. And from Faraday's law that tangential $\mathbf{E}$ field is continuous, then $\hat{n} \times \mathbf{E} = 0$ on the conductor surface since $\hat{n} \times \mathbf{E} = 0$ inside the conductor. Hence, the electric field has to be normal to the conductor surfaces. Figure 32.2 shows the static electric field, in the quiescent state, between two conductors (even though they are not PECs). Moreover, since $\mathbf{E} = 0$ inside the conductor, $\nabla\Phi = 0$ implying that the potential is a constant inside a conductor at equilibrium.

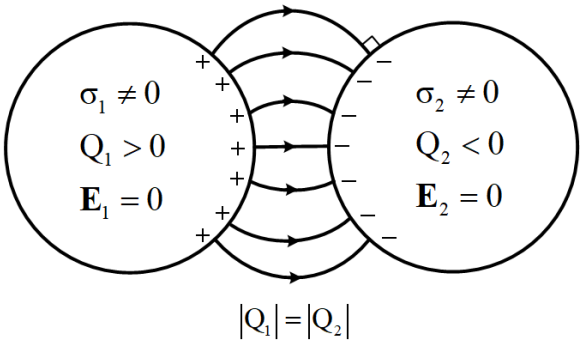


Figure 32.2: The objects can just be conductors, and in the quiescent state (static state), the tangential electric field will be zero on their surfaces. Also, $\mathbf{E} = 0$ inside the conductor, or $\nabla\Phi = 0$, or Φ is a constant inside.

32.1.2 Relaxation Time

The time it takes for the charges to move around until they reach their quiescent distribution such that $\mathbf{E} = 0$ is called the relaxation time. It is very much similar to the RC time constant of an RC circuit consisting of a resistor in series with a capacitor (see Figure 32.3). It can be proven that this relaxation time τ is related to ε/σ, but the proof is beyond the scope of this course at this point: it is worthwhile to note that this constant has the same unit as the RC time constant of an RC circuit where a charged capacitor relaxes as $\exp(-t/\tau)$ where the relaxation time $\tau = RC$. Note that when $\sigma \to \infty$, the relaxation time is zero. In other words, in a perfect conductor or a superconductor, the charges reorient themselves instantaneously if the external field is time-varying so that $\mathbf{E}(t) = 0$ always inside the conductor.[1]

[1]It is to be noted that a PEC is a fictitious medium and the closest to it is the superconductor.

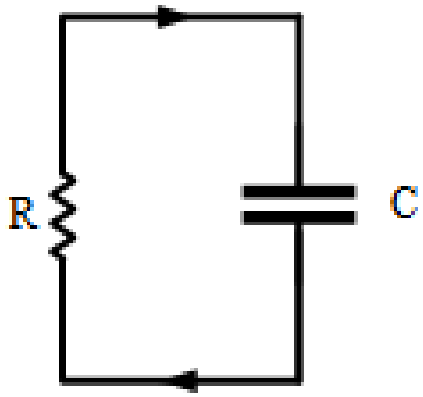


Figure 32.3: The relaxation (or disappearance of accumulated charges) in a conductive object is similar to the relaxation of charges from a charged capacitance in an RC circuit as shown.

Electrostatic shielding or low-frequency shielding is important at low frequencies. The Faraday cage or Faraday shield is an important application of such a shielding (see Figure 32.4). By grounding the Faraday cage, the potential inside the cage is set to zero[2] [231]. Earnshaw's theorem says that for Laplaces problems, the maximum or minimum of a region V bounded by the surface S has to be on the surface S. Thus, a volume V bounded by a surface S with equipotential, the potential inside the volume V has to be a constant. Thus the electric field $\mathbf{E} = -\nabla\Phi$ has to be zero.

[2]Whether if the potential is zero is immaterial, since potential is a relative concept. But in electrical engineering, it is customary to call the ground potential to be zero.

Figure 32.4: Faraday cage demonstration on volunteers in the Palais de la Découverte in Paris (courtesy of Wikipedia). When the cage is grounded, the potential at the surface of the cage is zero. By the solution to Laplace's equation, the potential inside the cage is a constant. Hence, the electric field inside the cage is zero. Charges will surge from the ground to the cage surface to ensure zero potential inside the cage. Therefore, a grounded Faraday cage effectively shields the external fields from entering the cage.

However, if the conductor charges are induced by an external electric field that is time varying, then the charges have to constantly redistribute/re-orient themselves to try to shield out the incident time-varying electric field. Currents have to be constantly flowing around the conductor. Then the electric field cannot be zero inside the conductors as shown in Figure 32.5. In other words, an object with finite conductivity cannot shield out completely a time-varying electric field. It can be shown that the depth of penetration of the field into the conductive object is about a skin depth $\delta = \sqrt{2/(\omega\mu\sigma)}$. Or the lower the frequency ω or the conductivity σ, the larger the penetration depth.

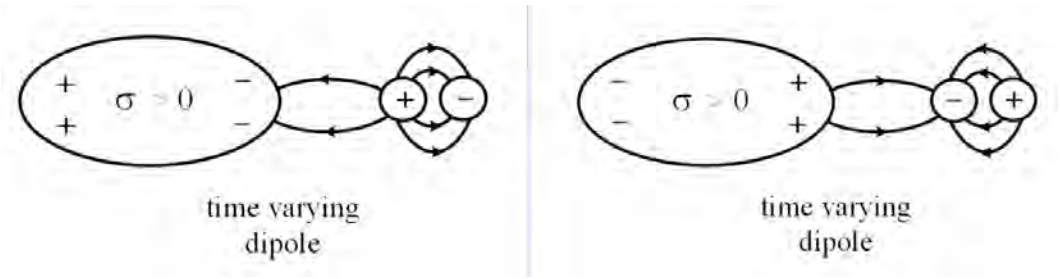


Figure 32.5: If the source that induces the charges on the conductor is time varying, the current in the conductor is always nonzero because the charges have to move around to respond to the external time-varying charges. The two figures above show the orientation of the charges for two snap-shots in time. In other words, a time-varying field can penetrate the conductor to approximately within a skin-depth $\delta = \sqrt{2/(\omega\mu\sigma)}$ from a plane-surface/plane-wave approximation.

For a perfect electric conductor (PEC), $\mathbf{E} = 0$ inside with the following argument: Because if $\mathbf{E} \neq 0$, then $\mathbf{J} = \sigma\mathbf{E}$ where $\sigma \to \infty$. Let us assume inside a PEC an infinitesimally small time-varying electric field in the PEC to begin with. It will induce an infinitely large electric current, and hence an infinitely large time-varying magnetic field. An infinite time-varying magnetic field in turn yields an infinite electric field that will drive an electric current, and these fields and current will be infinitely large. This is an unstable and escalating chain of events if it is true. Moreover, it will generate infinite energy in the system, which is not possible. Hence, the only physical possibility for a stable solution is for the time-varying electromagnetic fields to be zero inside a PEC.

Thus, for the PEC, the charges can re-orient themselves instantaneously on the surface when the inducing (incident or impinging) electric fields from outside are time varying. In other words, the relaxation time $\tau = \varepsilon/\sigma$ is zero. As a consequence, the time-varying electric field $\mathbf{E}$ is always zero inside PEC, with $\hat{n} \times \mathbf{E} = 0$ on the surface of the PEC, even for time-varying fields.

It is to be noted that when a point charge density is placed in a lossy conductive medium, closed form solution for the potential exists in the time-domain as well [108, p. 232-235]. In fact, $\tau = 0$ when $\sigma \to \infty$, implying that a point charge density disappears instantly when placed in a PEC medium. But this is an idealization. As we have seen from the Drude-Lorentz-Sommerfeld model, nothing is truly instantaneous in the real world!

32.2 Image Theory

The image theory here in electromagnetics is quite different from that in optics. As mentioned before, when the frequency of the fields is high, the waves associated with the fields can be described by rays. Therefore ray optics can be used to solve many high-frequency problems. We can use ray optics to understand how an image is generated in a mirror. But the image theory in electromagnetics is quite different from that in ray optics.

Image theory can be used to derived closed form solutions to boundary value problems when the geometry is simple with a lot of symmetry, especially for electrostatics and magnetostatics. These closed form solutions in turn offer physical insight into more complicated problems. This theory or theorem is also discussed in many textbooks [232, 46, 57, 1, 85, 68, 104].

32.2.1 Electric Charges and Electric Dipoles

Image theory for a flat conductor surface or a half-space is quite easy to derive. To see that, we can start with electro-static theory of putting a positive charge above a flat plane. As mentioned before, for electrostatics, the plane or half-space does not have to be a perfect conductor, but only a conductor (or a metal). From the previous Section 32.1.1, the tangential static electric field on the surface of the conductor has to be zero.

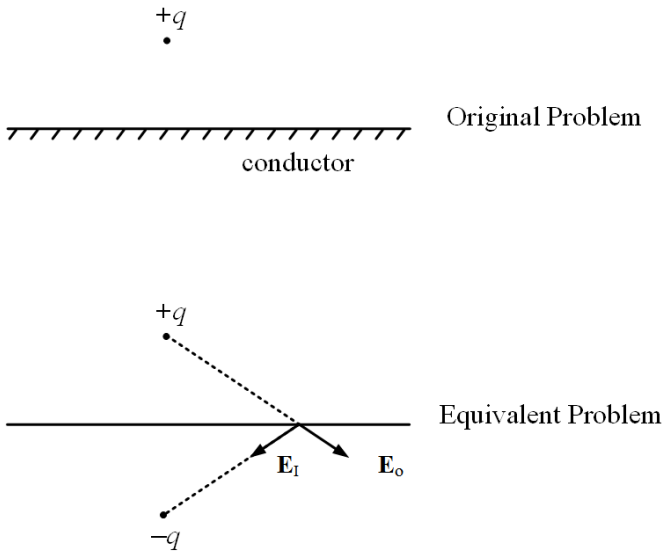


Figure 32.6: The use of image theory to solve the BVP (boundary value problem) of a point charge on top of a conductor. The boundary condition is that $\hat{n} \times \mathbf{E} = 0$ on the conductor surface. By placing a negative charge judiciously with respect to the original charge, by the principle of linear superposition, both of them produce a total field with no tangential component at the interface.

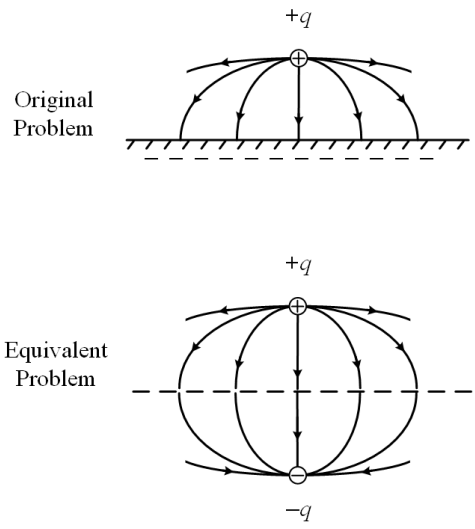


Figure 32.7: By image theory, the total electric field of the original problem and the equivalent problem when we add the total electric field due to the original charge and the image charge. It is to be noted that if we take the top half of the bottom figure, it is equivalent to the field of the top figure. The field for the bottom figure is easily obtained by superposing the fields of two opposite charges symmetrically located.

By the principle of linear superposition, the tangential static electric field can be canceled by putting an image charge of opposite sign at the symmetric mirror image location of the original charge. This is shown in Figure 32.6. Now we can mentally add the total field due to these two charges. When the total static electric field due to the original charge and image charge is sketched, it will look like that in Figure 32.7. It is seen that the static electric field satisfies the boundary condition that $\hat{n} \times \mathbf{E} = 0$ at the conductor interface due to symmetry.

An electric dipole is made from a positive charge placed in close proximity to a negative charge. Using the fact that an electric charge reflects to an electric charge of opposite polarity above a conductor, one can easily see that a static horizontal electric dipole reflects to a static horizontal electric dipole of opposite polarity. By the same token, a static vertical electric dipole reflects to static vertical electric dipole of the same polarity as shown in Figure 32.8.

If this electric dipole is a Hertzian dipole whose field is time-varying, then one needs a PEC surface to shield out the electric field. Also, the image charges will follow the original dipole charges instantaneously except for the retardation effect. Thus the image theory for static electric dipoles over a half-space still holds true if the dipoles now become Hertzian dipoles, but over a PEC surface.

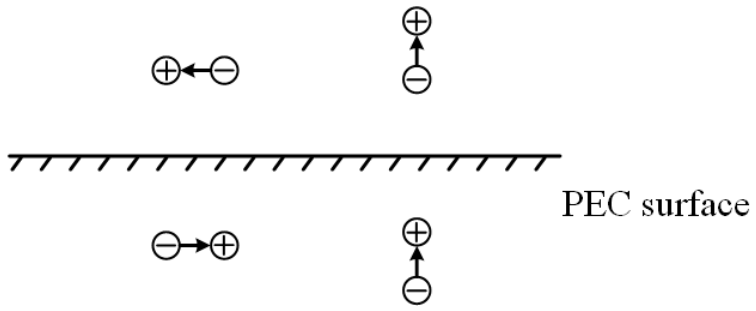


Figure 32.8: By image theory, on a conductor surface, a horizontal static dipole reflects to one of opposite polarity, while a static vertical dipole reflects to one of the same polarity. If the dipoles are time-varying, then a PEC will have a same reflection rule.

32.2.2 Magnetic Charges and Magnetic Dipoles

A static magnetic field can penetrate a conductive medium. This is apparent from our experience when we play with a bar magnet over a copper sheet: the magnetic field from the magnet can still be experienced by iron filings put on the other side of the copper sheet.

However, this is not the case for a time-varying magnetic field. Inside a conductive medium, a time-varying magnetic field will produce a time-varying electric field, which in turn produces the conduction current via $\mathbf{J} = \sigma\mathbf{E}$. This is termed eddy current, which by Lenz's law, produces an opposing magnetic field that repels the magnetic field from entering the conductive medium.[3]

Now, consider a static magnetic field penetrating into a perfect electric conductor, a minute amount of time variation will produce an electric field, which in turn produces an infinitely large eddy current. So the stable state for a static magnetic field inside a PEC is for it to be expelled from the perfect electric conductor. This in fact is what we observe when a magnetic field is brought near a superconductor. Therefore, for the static magnetic field, where $\mathbf{B} = 0$ inside the PEC, then $\hat{n} \cdot \mathbf{B} = 0$ on the PEC surface (see Figure 32.9).

[3]The repulsive force occurs by virtue of energy conservation. Since "work done" is needed to set the eddy current in motion in the conductor, or to impart kinetic energy to the electrons forming the eddy current, a repulsive force is felt in Lenz's law so that work is done in pushing the magnetic field into the conductive medium.

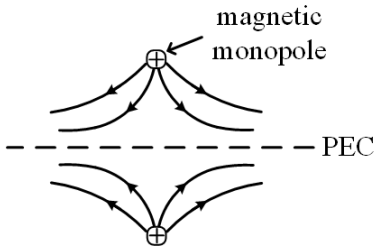


Figure 32.9: On a PEC surface, the requisite boundary condition is $\hat{n} \cdot \mathbf{B} = 0$. Hence, a static magnetic monopole on top of a PEC surface will have magnetic field distributed as shown. By image theory, such a distribution of the $\mathbf{B}$ field can be obtained by adding a magnetic monopole of the same polarity at its image point.

Now, assuming that a magnetic monopole exists, it will reflect to itself on a PEC surface so that $\hat{n} \cdot \mathbf{B} = 0$ as shown in Figure 32.9. Therefore, a magnetic charge reflects to a charge of similar polarity on the PEC surface.

By extrapolating this to magnetic dipoles, they will reflect themselves to the magnetic dipoles as shown in Figure 32.10. A horizontal magnetic dipole reflects to a horizontal magnetic dipole of the same polarity, and a vertical magnetic dipole reflects to a vertical magnetic dipole of opposite polarity. Then, a vertical dipolar bar magnet near a superconducting hals-space reflects to a vertical bar magnet of opposite polarity: it can be levitated by a superconductor half-space when this magnet is placed close to it. This is also known as the Meissner effect [233], which is shown in Figure 32.11.

A time-varying magnetic dipole can be made from a electric current loop. Over a PEC, a time-varying magnetic dipole will reflect the same way as a static magnetic dipole as shown in Figure 32.10.

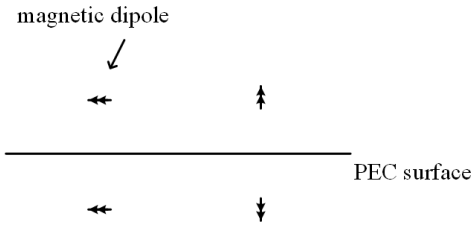


Figure 32.10: Using the rule of how magnetic monopole reflects itself on a PEC surface, the reflection rules for magnetic dipoles can be ascertained. (Magnetic dipoles are often denoted by double arrows.)

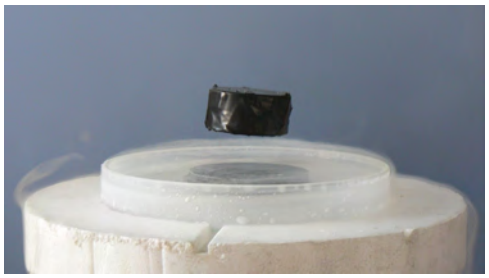


Figure 32.11: On a PEC (superconducting) surface, a static vertical magnetic dipole above it (formed by a small permanent bar magnet here) reflects to an image of a bar magnet of opposite polarity. Hence, the magnetic dipoles repel each other displaying the Meissner effect. The magnet, because of the repulsive force from its image, levitates above the superconductor (courtesy of Wikipedia [234]).

32.2.3 Perfect Magnetic Conductor (PMC) Surfaces

Magnetic conductor does not come naturally in this world since there are no free-moving magnetic charges around. Magnetic monopoles are yet to be discovered. On a PMC surface, by duality, $\hat{n} \times \mathbf{H} = 0$. At low frequency, it can be mimicked by a high μ material. One can see that for magnetostatics, at the interface of a high μ material and air, the magnetic flux is approximately normal to the surface, resembling the $\mathbf{H}$ field near a PMC surface (see Figure 32.12.

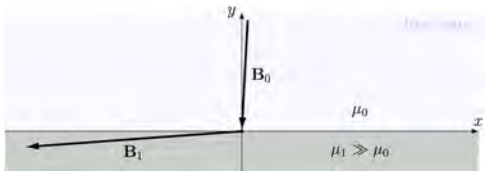


Figure 32.12: At the interface between free space and a high-μ material, one can see that magnetic field in the free-space region is almost normal to the surface in order for normal $\mathbf{B}$ and tangential $\mathbf{H}$ to be continuous. Hence, looking from the free-space side, the high-μ material resembles a PMC.

High μ materials are hard to find at higher frequencies. Since $\hat{n} \times \mathbf{H} = 0$ on such a surface, no electric current can flow on such a surface. Hence, a PMC can be mimicked by a surface where no surface electric current can flow. This has been achieved in microwave engineering with a mushroom surface as shown in Figure 32.13 [235]. The mushroom structure consisting of a wire and an end-cap, can be thought of as forming an LC tank circuit. Close to the resonance frequency of this tank circuit, the surface of mushroom structures essentially becomes open circuits with no or little current flowing on the surface, or $\mathbf{J}_s \cong 0$. In other words, $\hat{n} \times \mathbf{H} \cong 0$. This resembles a PMC surface, because with no surface electric current can flow on such a surface, the tangential

magnetic field is small, the hallmark of a good magnetic conductor, by using the duality principle.

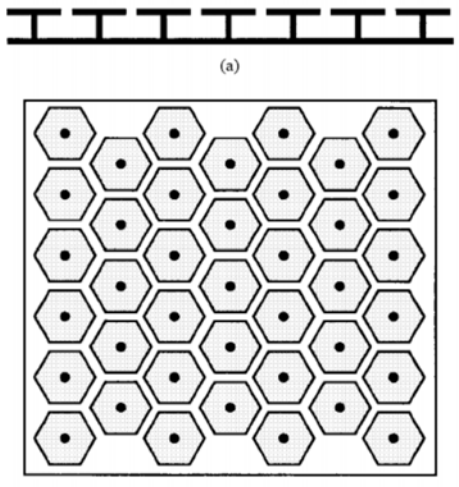


Figure 32.13: A mushroom structure operates like an LC tank circuit. At the right resonant frequency, the surface resembles an open-circuit surface where no current can flow. Hence, tangential magnetic field is zero resembling a perfect magnetic conductor (courtesy of Sievenpiper [235]).

Mathematically, a surface that is dual to the PEC surface is the perfect magnetic conductor (PMC) surface. The magnetic dipole is also dual to the electric dipole. Thus, over a PMC surface, these electric and magnetic dipoles will reflect differently as shown in Figure 32.14. One can go through Gedanken experiments and verify that the reflection rules are as shown in the figure.

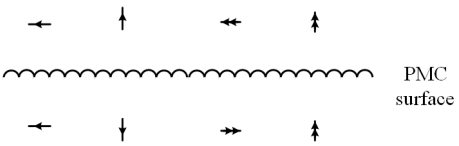


Figure 32.14: Reflection rules for electric and magnetic dipoles over a PMC surface. Magnetic dipoles are denoted by double arrows.

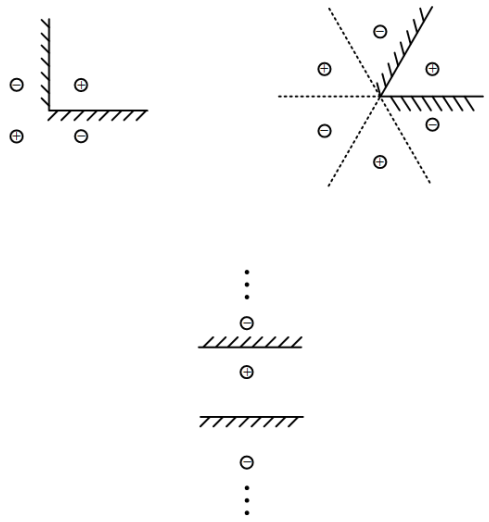


Figure 32.15: Image theory for multiple images [33].

32.2.4 Multiple Images

For the geometry shown in Figure 32.15, one can start with electrostatic theory, and convince oneself that $\hat{n} \times \mathbf{E} = 0$ on the metal surface with the placement of charges as shown. For conducting media, the charges will relax to the quiescent distribution after the relaxation time. For PEC surfaces, one can extend these cases to time-varying dipoles because the charges in the PEC medium can re-orient instantaneously (i.e. with zero relaxation time) to shield out or expel the **E** and **H** fields. Again, one can repeat the above exercise for magnetic charges, magnetic dipoles, and PMC surfaces.

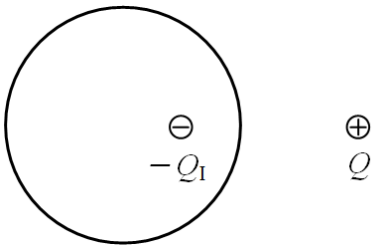


Figure 32.16: Image theory for a point charge near a cylinder or a sphere can be found in closed form. Details are given in [33].

32.2.5 Some Special Cases—Spheres, Cylinders, and Dielectric Interfaces

One curious case is for a static charge placed near a conductive sphere (or cylinder) as shown in Figure 32.16.[4] A charge of $+Q$ reflects to a charge of $-Q_I$ inside the sphere where $Q_I \neq Q$. For electrostatics, the sphere (or cylinder) need only be a conductor. However, this cannot be generalized to electrodynamics or a time-varying problem, because of the retardation effect: A time-varying dipole or charge will be felt at different points asymmetrically on the surface of the sphere from the original and image charges. Exact cancelation of the tangential electric field on the surface of the sphere or cylinder cannot occur for time-varying field.

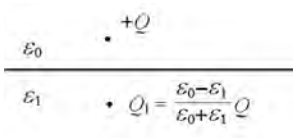


Figure 32.17: A static charge over a dielectric interface can be found in closed form using Fourier transform technique to be discussed later. The solution is beyond the scope of this chapter at this point.

When a static charge is placed over a dielectric interface, image theory can be used to find the closed form solution. This solution can be derived using Fourier transform technique which we shall learn later [37]. It can also be extended to multiple interfaces. But image theory cannot be used in a simple manner for the electrodynamic case due to the different speed of light in different media, giving rise to different retardation effects.

[4] This is worked out in detail in p. 48 and p. 49, Ramo et al [33].

Exercises for Lecture 32

Problem 32-1:

(i) Explain why in the case of a conductor with finite conductivity, the electric field is completely shielded from the inside of the conductor when the electric field is static.

(ii) Explain why a conductor with finite conductivity cannot shield out a magnetic field in the static limit.

(iii) Explain why a bar magnet can levitate on top of a superconductor.

(iv) Using image theorem, explain why the boundary conditions are satisfied in the examples shown in Figure 32.15.

(v) By using the principle of linear superposition, for Figure 32.17, show that by image theory that the normal flux and tangential field are continuous when they are produced by the original source plus its image source.

Chapter 33

High Frequency Solutions, Gaussian Beams

When the frequency is very high, the wavelength of electromagnetic wave becomes very short. In this limit, many solutions to Maxwell's equations can be found approximately. These solutions offer a very different physical picture of electromagnetic waves, and they are often used in optics where the wavelength is short. So it was no surprise that for a while, optical fields were thought to satisfy a very different set of equations from those of electricity and magnetism. Therefore, it came as a surprise that when it was later revealed that in fact, optical fields satisfy the same Maxwell's equations as the fields from electricity and magnetism!

In this lecture, we shall seek approximate solutions to Maxwell's equations or the wave equations when the frequency is high or the wavelength is short compared to the geometry that the wave interacts with. High frequency approximate solutions are important in many real-world applications. This is possible when the wavelength is much smaller than the size of the structure. This can occur even in the microwave regime where the wavelength is not that small, but much smaller than the size of the structure. This is the case when microwave interacts with reflector antennas for instance. It is also the transition from waves regime to the optics (or ray optics) regime in the solutions of Maxwell's equations. Often times, the term "quasi-optical" is used to describe the solutions in this regime.

In the high frequency regime, or when we are far away from a source much larger than the Rayleigh distance (see Section 30.3.1), the field emanating from a source resembles a spherical wave. Moreover, when the wavelength is much smaller than the radius of curvature of the wavefront, the spherical wave can be approximated by a local plane wave. Thus we can imagine rays to be emanating from a finite source forming the spherical wave. The spherical wave will ultimately be approximated by plane waves locally at the observation point.[1] This will simplify the solutions in many instances. For instance, ray tracing can be used to track how these rays can propagate, bounce, or "ricochet" in a complex environment. In fact, it is now done in a movie industry to give "realism" to simulate the nuances of how light rays bounce around in a room, and reflect off objects.

[1] We shall learn later that a ray can be approximated by a bundle of plane waves almost parallel to each other.

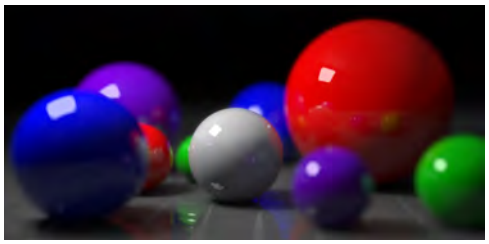


Figure 33.1: Ray-tracing technique can be used in the movie industry to produce realism in synthetic images (courtesy of Wikipedia).

33.1 Tangent Plane Approximations

We have learnt that reflection and transmission of waves at an infinitely large flat surface can be solved in closed form. The important point here is the physics of phase matching. Due to phase matching, we have derived the law of Fresnel reflection, transmission and Snell's law [61].[2]

When a surface is not flat anymore, there is no closed form solution. But when a surface is curved such that the radius of curvature is much larger than the wavelength, an approximate solution can be found. This is obtained by using a *local tangent-plane approximation*, which is a good approximation when the frequency is high or the wavelength is short. It is similar in spirit that we can approximate a spherical wave by a local plane wave at the spherical wave front when the wavelength is short compared to the radius of curvature of the wavefront.

When the wavelength is short, phase matching happens locally, and the Fresnel law of reflection, transmission, and Snell's law are satisfied approximately as shown in Figure 33.2. The tangent plane approximation is the basis for the geometrical optics (GO) approximation [237, 34]. In GO, light waves are replaced by light rays. As mentioned before, a light ray is a part of a spherical wave where locally, the spherical wave can be approximated by a plane wave. The reflection and transmission of these rays at an interface is then estimated using the local tangent plane approximation and local Fresnel reflection and transmission coefficients. This is also the basis for high frequency solutions where lens, ray optics, or ray tracing can be used. From it, lens technology is derived (see Figure 33.3). [238, 239].[3]

Many real world problems do not have closed-form solutions, and have to be treated with approximate methods. In addition to geometrical optics approximations mentioned above, asymptotic methods are also used to find approximate solutions. Asymptotic methods imply finding a solution when there is a large parameter in the problem. In this case, it is usually the frequency. Such high-frequency approximate methods are discussed in [240, 241, 242, 243, 244].

[2]This law is also known in the Islamic world in A.D. 984 [236].

[3]Please note that the tangent plane approximation is invalid near a sharp corner or an edge. The solution has to be augmented by additional diffracted wave coming from the edge or the corner.

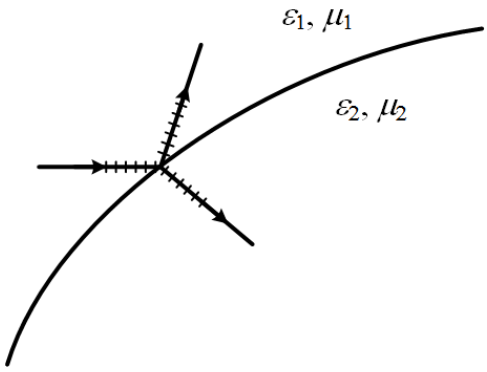


Figure 33.2: In the tangent plane approximation, the surface where reflection and refraction occur is assumed to be locally flat. Thus, phase-matching is approximately satisfied, yielding locally the law of reflection, transmittion, and Snell's law.

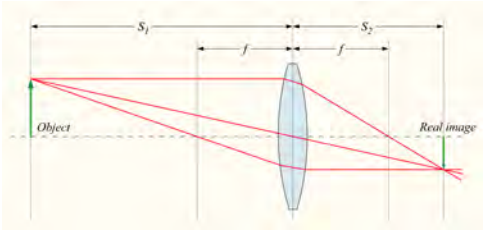


Figure 33.3: Tangent plane approximations can also be made at dielectric interfaces so that Fresnel reflection and transmission coeffients can be used to ascertain the interaction of light rays with a lens. Also, one can use ray tracing to understand the working of an optical lens (courtesy of Wikipedia).

33.2 Fermat's Principle

Fermat's principle (1600s) [245, 61] says that a light ray follows the path that takes the shortest time delay between two points.[4] Since time delay is related to the phase shift, and that a light ray can be locally approximated by a plane wave, this can be re-stated that a plane wave follows the path that has a minimal phase shift. This principle can be used to derive the law of reflection, transmission, and refraction for light rays. It can be used as the guiding principle for ray tracing as well.

[4]This eventually give rise to the principle of least action, which is a wonderful gift of Nature! Nature finds the simplest and most efficient solution in the real world.

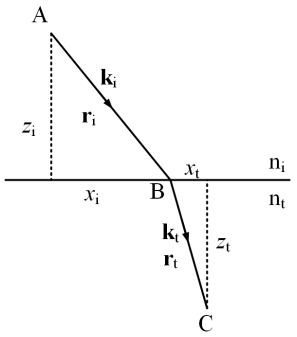


Figure 33.4: In Fermat's principle, a light ray, when propagating from point A to point C, takes the path of least time delay in the time domain, and hence, least phase shift in the frequency domain.

Assume two points A and C are in two different half spaces as shown in Figure 33.4. Then the phase delay between the two points, per Figure 33.4, can be written as[5]

$$P = \mathbf{k}_i \cdot \mathbf{r}_i + \mathbf{k}_t \cdot \mathbf{r}_t \tag{33.2.1}$$

In the above, $\mathbf{k}_i$ is parallel to $\mathbf{r}_i$, so is $\mathbf{k}_t$ is parallel to $\mathbf{r}_t$. As this is the shortest path with minimum phase shift or time delay, according to Fermat's principle, another other path will be longer giving rise to more phase shift. In other words, if B were to move slightly to another point, a longer path with more phase shift or time delay will ensue, or that B is the stationary point for the path length or phase shift when the location of B is changed.

Specializing (33.2.1) to a 2D picture, then the phase shift as a function of x_i is stationary. This is shown in Figure 33.4, we have $x_i + x_t = \text{const}$. Therefore, taking the derivative of (33.2.1) or the phase change with respect to x_i, assuming that $\mathbf{k}_i$ and $\mathbf{k}_t$ do not change as B is moved slightly,[6]

$$\frac{\partial P}{\partial x_i} = 0 = k_{ix} - k_{tx} \tag{33.2.2}$$

The above yields the law of refraction that $k_{ix} = k_{tx}$, which is just Snell's law; it can also be obtained by phase matching as we have shown earlier. This law was also known in the Islamic world to Ibn Sahl in A.D. 984 [236].

[5]In this course, for wavenumber, we use k and β interchangeably, where k is prevalent in optics and β is used in microwaves.

[6]One can show that as the separations between A, B, and C are large, and if the change in x_i is Δx_i, the changes in $\mathbf{k}_i$ and $\mathbf{k}_t$ are small. The change in phase shift mainly comes from the change in x_i. Alternatively, we can write $P = k_i r_i + k_t r_t$, and let $r_i = \sqrt{x_i^2 + z_i^2}$, and $r_t = \sqrt{x_t^2 + z_t^2}$, and take the derivative with respect to x_i, one would also get the same answer.

33.2.1 Generalized Snell's Law

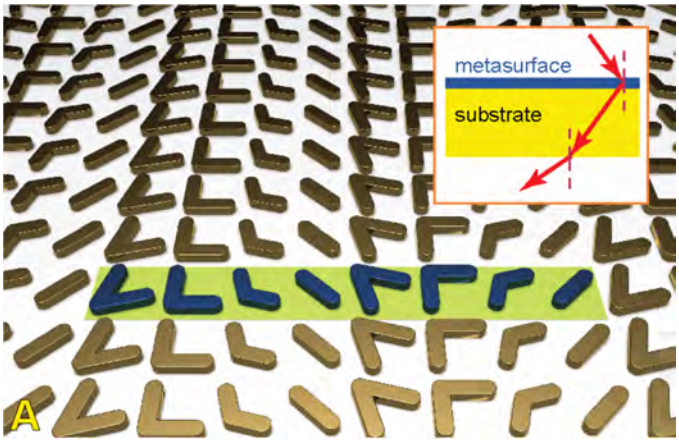


Figure 33.5: A phase screen which is position dependent can be made using nano-fabrication and designed with commercial software for solving Maxwell's equations. In such a case, one can derive a generalized Snell's law to describe the diffraction of a wave by such a surface (courtesy of Capasso's group [246]).

Metasurfaces are prevalent these days due to advances in nano-fabrication and numerical simulation. One of them is shown in Figure 33.5. Such a metasurface can be thought of as a phase screen, providing additional phase shift for the light as it passes through it. Moreover, the added phase shift can be controlled to be a function of position because of advances in nano-fabrication technology and commercial software for numerical simulation.

To model this phase screen, we add an additional function $\Phi(x, y)$ to (33.2.1), namely that

$$P = \mathbf{k}_i \cdot \mathbf{r}_i + \mathbf{k}_t \cdot \mathbf{r}_t - \Phi(x_i, y_i) \tag{33.2.3}$$

Now applying Fermat's principle that there should be minimal phase delay, and taking the derivative of the above with respect to x_i, one gets

$$\frac{\partial P}{\partial x_i} = k_{ix} - k_{tx} - \frac{\partial \Phi(x_i, y_i)}{\partial x_i} = 0 \tag{33.2.4}$$

The above yields that the generalized Snell's law [246] that

$$k_{ix} - k_{tx} = \frac{\partial \Phi(x_i, y_i)}{\partial x_i} \tag{33.2.5}$$

It implies that the transmitted light can be directed to other angles due to the additional phase screen.[7]

[7]Such research is also being pursued by V. Shalaev and A. Boltasseva's group at Purdue U [247].

33.3 Gaussian Beam

At this point, we will take a departure to study another high-frequency phenomenon, that of the Gaussian beam. We have seen previously that in a source-free medium, using vector and scalar potential formulation, we arrive at simpler Helmholtz wave equations for $\mathbf{A}$ and Φ, namely,

$$\nabla^2 \mathbf{A} + \omega^2 \mu\varepsilon \mathbf{A} = 0 \tag{33.3.1}$$

$$\nabla^2 \Phi + \omega^2 \mu\varepsilon \Phi = 0 \tag{33.3.2}$$

The above are four scalar equations; and the Lorenz gauge

$$\nabla \cdot \mathbf{A} = -j\omega\mu\varepsilon\Phi \tag{33.3.3}$$

connects $\mathbf{A}$ and Φ. We can examine the solution of $\mathbf{A}$ such that

$$\mathbf{A}(\mathbf{r}) = \mathbf{A}_0(\mathbf{r}) e^{-j\beta z} \tag{33.3.4}$$

where $\mathbf{A}_0(\mathbf{r})$ is a slowly varying function while $e^{-j\beta z}$ is rapidly varying in the z direction. (Here, $\beta = \omega\sqrt{\mu\epsilon}$ is the wavenumber.) This is primarily a quasi-plane wave propagating predominantly in the z-direction. We know this to be the case in the far field of a source, but let us assume that this form persists less than the far field, namely, in the Fresnel zone as well. Taking the x component of (33.3.4), we have[8]

$$A_x(\mathbf{r}) = \Psi(\mathbf{r}) e^{-j\beta z} \tag{33.3.5}$$

where $\Psi(\mathbf{r}) = \Psi(x, y, z)$ is a scalar slowly varying envelope function of x, y, and z, whereas $e^{-j\beta z}$ is a rapidly varying function of z when β is large or the frequency is high.

33.3.1 Derivation of the Paraxial/Parabolic Wave Equation

Substituting (33.3.5) into (33.3.1), and taking the double z derivative first in the Laplacian operator ∇^2, we arrive at

$$\frac{\partial^2}{\partial z^2}\left[\underbrace{\Psi(x,y,z)}_{\text{slow}}\underbrace{e^{-j\beta z}}_{\text{fast}}\right] = \left[\frac{\partial^2}{\partial z^2}\Psi(x,y,z) - 2j\beta\frac{\partial}{\partial z}\Psi(x,y,z) - \beta^2\Psi(x,y,z)\right]e^{-j\beta z} \tag{33.3.6}$$

Consequently, after substituting the above into the x component of (33.3.1), making use of the definition of ∇^2, we obtain an equation for $\Psi(\mathbf{r})$, the slowly varying envelope as

$$\frac{\partial^2}{\partial x^2}\Psi + \frac{\partial^2}{\partial y^2}\Psi - 2j\beta\frac{\partial}{\partial z}\Psi + \frac{\partial^2}{\partial z^2}\Psi = 0 \tag{33.3.7}$$

where the last term of (33.3.6) containing β^2 on the right-hand side cancels with the term coming from $\omega^2\mu\varepsilon\mathbf{A}$ of (33.3.1).

[8] Also, the wave becomes a transverse wave in the far field, and keeping the transverse component suffices.

So far, no approximation has been made in the above equation. Since β is linearly proportional to frequency ω, when $\beta \to \infty$, or in the high frequency limit,

$$\left| 2j\beta \frac{\partial}{\partial z} \Psi \right| \gg \left| \frac{\partial^2}{\partial z^2} \Psi \right| \tag{33.3.8}$$

where we have assumed that Ψ is a slowly varying function of z within the lengthscale of a wavelength, such that $\beta\Psi \gg \partial/\partial z\Psi$, which in turn implies the above. In other words, (33.3.7) can be approximated by

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} - 2j\beta \frac{\partial \Psi}{\partial z} \approx 0 \tag{33.3.9}$$

The above is called the paraxial wave equation. It is also called the parabolic wave equation.[9] It implies that the $\boldsymbol{\beta}$ vector of the wave is approximately parallel to the z axis, or $\beta_z \cong \beta$ to be much greater than β_x and β_y, and hence, the name.

33.3.2 Finding a Closed Form Solution

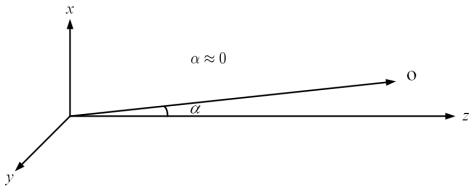


Figure 33.6: This figure shows when the paraxial approximation can be made: when the field is being observed very close to the z axis and far away, the wave physics is different.

A closed form solution to the paraxial wave equation can be obtained by a simple mathematical trick.[10] It is known that

$$A_x(\mathbf{r}) = \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{33.3.10}$$

is the exact solution to

$$\nabla^2 A_x + \beta^2 A_x = 0 \tag{33.3.11}$$

as long as $\mathbf{r} \neq \mathbf{r}'$. One way to ensure that $\mathbf{r} \neq \mathbf{r}'$ always is to let $\mathbf{r}' = -\hat{z}jb$, a complex number. Then (33.3.10) is always a solution to (33.3.11) for all $\mathbf{r}$, because $|\mathbf{r} - \mathbf{r}'| \neq 0$ always since $\mathbf{r}'$

[9]The paraxial wave equation, the diffusion equation and the Schrodinger equation are all classified as parabolic equations in mathematical parlance [248, 52, 249, 37].

[10]Introduced by Georges A. Deschamps of UIUC [250].

is complex. Then, we should next make a paraxial approximation to the solution (33.3.10) by assuming that $x^2 + y^2 \ll z^2$. By so doing, it follows that

$$\begin{aligned} |\mathbf{r} - \mathbf{r}'| &= \sqrt{x^2 + y^2 + (z + jb)^2} \\ &= (z + jb)\left[1 + \frac{x^2 + y^2}{(z + jb)^2}\right]^{1/2} \\ &\approx (z + jb) + \frac{x^2 + y^2}{2(z + jb)} + \ldots, \qquad |z + jb| \to \infty \end{aligned} \tag{33.3.12}$$

where Taylor series has been used in approximating the last term. And then using the above approximation in (33.3.10) yields

$$A_x(\mathbf{r}) \approx \frac{e^{-j\beta(z+jb)}}{4\pi(z + jb)} e^{-j\beta \frac{x^2+y^2}{2(z+jb)}} \approx e^{-j\beta z}\Psi(\mathbf{r}) \tag{33.3.13}$$

Notice that again, the second term in (33.3.12) is ignored in the denominator, but not in the exponent. By comparing the above with (33.3.5), we can identify

$$\Psi(x, y, z) \cong A_0 \frac{jb}{z + jb} e^{-j\beta \frac{x^2+y^2}{2(z+jb)}} \tag{33.3.14}$$

where A_0 is used to absorbed the constants to simplify the expression and make the rest of the expression dimensionless. By separating the exponential part into the real part and the imaginary part, viz.,

$$\frac{x^2 + y^2}{2(z + jb)} = \frac{x^2 + y^2}{2}\left(\frac{z}{z^2 + b^2} - j\frac{b}{z^2 + b^2}\right) \tag{33.3.15}$$

and writing the prefactor in terms of amplitude and phase gives,

$$\frac{jb}{z + jb} = \frac{1}{\sqrt{1 + z^2/b^2}} e^{j \tan^{-1}(\frac{z}{b})} \tag{33.3.16}$$

We then have

$$\Psi(x, y, z) \cong \frac{A_0}{\sqrt{1 + z^2/b^2}} e^{j \tan^{-1}(\frac{z}{b})} e^{-j\beta \frac{x^2+y^2}{2(z^2+b^2)} z} e^{-b\beta \frac{x^2+y^2}{2(z^2+b^2)}} \tag{33.3.17}$$

The above can be rewritten more suggestively as

$$\Psi(x, y, z) \cong \frac{A_0}{\sqrt{1 + z^2/b^2}} e^{-j\beta \frac{x^2+y^2}{2R}} e^{-\frac{x^2+y^2}{w^2}} e^{j\Psi} \tag{33.3.18}$$

where A_0 is a new constant introduced to absorb undesirable constants arising out of the algebra. In the above,

$$w^2 = \frac{2b}{\beta}\left(1 + \frac{z^2}{b^2}\right), \qquad R = \frac{z^2 + b^2}{z}, \qquad \Psi = \tan^{-1}\left(\frac{z}{b}\right) \tag{33.3.19}$$

For a fixed z, the parameters w, R, and Ψ are all constants. It is seen that the beam is Gaussian tapered in the x and y directions, and hence, the name Gaussian beam. Here, w is the beam waist which varies with z, and it is smallest when $z = 0$, or $w = w_0 = \sqrt{\frac{2b}{\beta}}$ at its starting value.

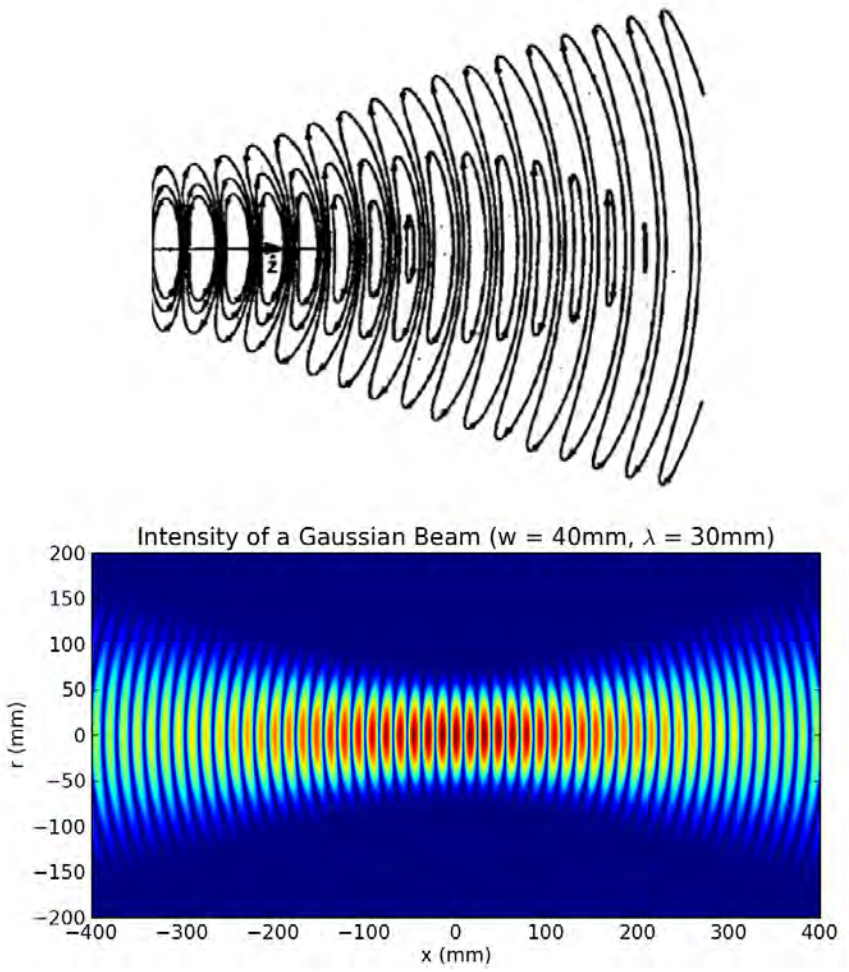


Figure 33.7: Electric field of a Gaussian beam in the x-z plane frozen in time. The wave moves to the right as time increases; here, $b/\lambda = 10/6$ (courtesy of Haus, Electromagnetic Noise and Quantum Optical Measurements [89]). The narrowest beam waist is given by $w_0/\lambda = \sqrt{b/(\lambda\pi)}$. The lower figure is a more colorful one (courtesy of Wikipedia). Notice that these are homogeneous solutions to the wave equation since the right-hand side is always zero.

And the term $\exp(-j\beta\frac{x^2+y^2}{2R})$ resembles the phase front of a spherical wave where R is its radius of curvature. This can be appreciated by studying a spherical wave front $e^{-j\beta R}$, and make a paraxial wave approximation, namely, letting $x^2+y^2 \ll z^2$ to get

$$\begin{aligned} e^{-j\beta R} = e^{-j\beta(x^2+y^2+z^2)^{1/2}} &= e^{-j\beta z\left(1+\frac{x^2+y^2}{z^2}\right)^{1/2}} \\ &\approx e^{-j\beta z - j\beta\frac{x^2+y^2}{2z}} \approx e^{-j\beta z - j\beta\frac{x^2+y^2}{2R}} \end{aligned} \tag{33.3.20}$$

In the last approximation, we assume that $z \approx R$ in the paraxial approximation. We see that the phase of the above wave field, minus the $-j\beta z$ term, is similar in form to the first phase term in (33.3.18). Hence, R in (33.3.18) can be thought of as the radius of curvature of the phase front or wave front.

The phase Ψ defined in (33.3.19) changes linearly with z for small z, and saturates to a constant for large z. This underscores the fact that $\Psi(\mathbf{r})$ is a slowly varying function, and also, the phase of the entire wave is due to the $\exp(-j\beta z)$ in (33.3.13) which is rapidly varying when β is large or the frequency is high. A cross section of the electric field due to a Gaussian beam is shown in Figure 33.7.

33.3.3 Other solutions

In general, the paraxial wave equation in (33.3.9) is of the same form as the Schrödinger equation which is of utmost importance in quantum theory. In recent years, the solution of this equation has made use of spill-over knowledge and terms from quantum theory, such as spin angular momentum (SAM) or orbital angular momentum (OAM) even though we are actually in the classical regime. But it is a partial differential equation which can be solved by the separation of variables just like the Helmholtz wave equation. (Hurrah to the power of separation of variables!) Therefore, in general, it has solutions of the form[11]

$$\Psi_{nm}(x,y,z) \sim \frac{1}{w} e^{-(x^2+y^2)/w^2} e^{-j\frac{\beta}{2R}(x^2+y^2)} e^{j(m+n+1)\tan^{-1}\frac{z}{b}} H_n\left(x\sqrt{2}/w\right) H_m\left(y\sqrt{2}/w\right) \tag{33.3.21}$$

where $H_n(\xi)$ is a Hermite polynomial of order n. Alternatively, the solutions can also be expressed in terms of Laguere polynomials , namely,

$$\Psi_{nm}(x,y,z) \sim \frac{1}{w} e^{-j\frac{\beta}{2R}\rho^2} - e^{-\rho^2/w^2} e^{+j(n+m+1)\tan^{-1}\left(\frac{z}{b}\right)} e^{jl\phi} \left(\frac{\sqrt{2}\rho}{w}\right) L^{n-m}_{\min(n,m)}\left(\frac{2\rho^2}{w^2}\right) \tag{33.3.22}$$

where $L^k_n(\xi)$ is the associated Laguerre polynomial. One can also generate a field due to a multipole source. When the source point is put into the complex space, then a complex Gaussian beam can be generated.

These Gaussian beams have rekindled recent excitement in the community because, in addition to carrying spin angular momentum as in a plane wave, they can carry orbital angular momentum

[11] See F. Pampaloni and J. Enderlein [251]. The author also thanks Bo ZHU for pointing errors in the earlier versions of these equations.

due to the complex transverse field distribution of the beams.[12] They harbor potential for optical communications as well as optical tweezers to manipulate trapped nano-particles. Figure 33.8 shows some examples of the cross section (xy plane) field plots for some of these beams. They are richly endowed with patterns implying that they can be used to encode information. These lights are also called structured lights [252].

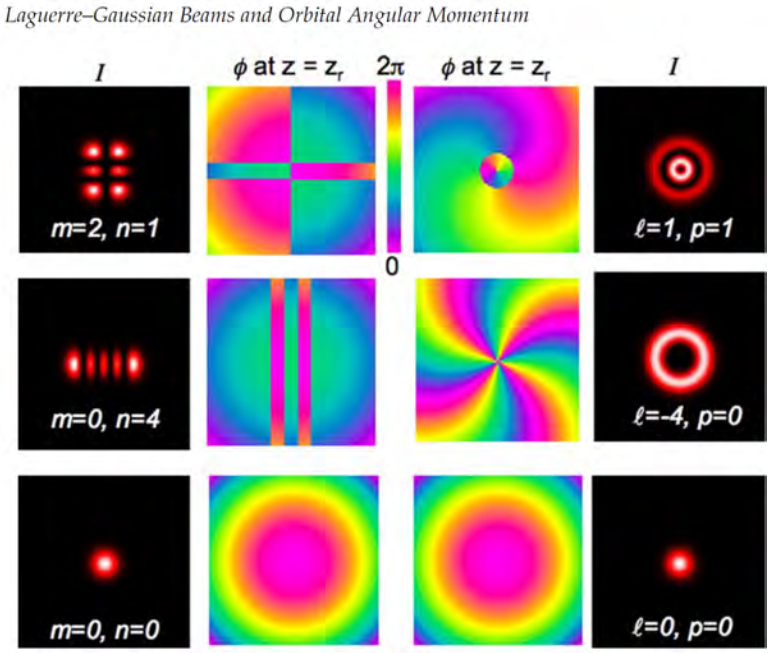


Figure 1.1 Examples of the intensity and phase structures of Hermite–Gaussian modes (*left*) an Laguerre–Gaussian modes (*right*), plotted at a distance from the beam waist equal to the Rayleig range. See color insert.

Figure 33.8: Examples of structured light. It can be used in encoding more information in optical communications (courtesy of L. Allen and M. Padgett's chapter in J.L. Andrew's book on structured light [252]).

[12] See D.L. Andrew, Structured Light and Its Applications and articles therein [252].

Exercises for Lecture 33

Problem 33-1:

(i) Verify the statement on Footnote 6 in Chapter 33 of the lecture book.

(ii) Explain why (33.3.8) is true in the high-frequency and paraxial limit.

(iii) Derive (33.3.18) and the expressions in (33.3.19) and give physical meanings to (33.3.18). Explain why R is the radius of curvature of the wavefront.

Chapter 34

Scattering of Electromagnetic Field

The scattering of electromagnetic field is an important and fascinating topic. It especially enriches our understanding of the interaction of light wave with matter. The wavelength of visible light is several hundred nanometers, with atoms and molecules ranging from nano-meters onward, light-matter interaction is richly endowed with interesting physical phenomena! Moreover, the myriads of hue generated by visible light has a great retina impact.

A source radiates a field, and ultimately, in the far field of the source, the field resembles a spherical wave which in turn resembles a plane wave. When a plane wave impinges on an object or a scatterer, the energy carried by the plane wave is deflected to other directions which is the process of scattering. This is akin to the physics of mode conversion at waveguide junctions that we have alluded to in Section 24.3.3: a pure single mode sprays into many other modes in order to satisfy the boundary condition. In the optical regime, the scattered light allows us to see objects, as well as admire all the beautiful hues and colors. In microwave, the scatterers cause the loss of energy carried by a plane wave.

A proper understanding of scattering theory reveals many physical phenomena around us. We will begin by studying Rayleigh scattering, which is scattering by small objects (or particles) compared to wavelength. With Rayleigh scattering of a simple sphere, we gain physical insight to many phenomena, such as why the sky is blue and the sunset is red!

34.1 Rayleigh Scattering

Rayleigh scattering is a solution to the scattering of light by small particles. These particles are assumed to be much smaller than the wavelength of light. The size of water molecule is about 0.25 nm, while the wavelength of blue light is about 450 nm. Since the particle size is much smaller than the wavelength, we can use quasi-static approximation to find a simple solution in the vicinity

of the small particle.[1]

This simple scattering solution offers us insight into nature of light scattering (see Figure 34.1). For instance, it explains why the sunset so magnificently beautiful, how birds and insects can navigate themselves without the help of a compass. By the same token, it also explains why in ancient times, the Vikings, as a seafaring people, could cross the Atlantic Ocean over to Iceland without the help of a magnetic compass as the Chinese did.

Figure 34.1: The magnificent beauty of nature can be partly explained by Rayleigh scattering [253, 186].

When a ray of light impinges on an object, we model the incident light as a plane electromagnetic wave (see Figure 34.2). The time-varying incident field polarizes the particle, making it into a small time-varying dipole, and it re-radiates like a Hertzian dipole. This is the gist of a scattering process: an incident field induces current (in this case, polarization current) on the scatterer. With the induced current, the scatterer re-radiates (or scatters). Without loss of generality, we can assume that the electromagnetic wave is polarized in the z direction and propagating in the x direction. We assume that the particle to be a small spherical particle with permittivity ε_s and radius a. Essentially, the particle sees a constant field as the plane wave impinges on it: Or in a word, the particle feels a quasi-electrostatic field in the incident field.

[1]Since the above is derived with phasor technique, it is also valid for complex permittivity and permeability as long as the wavelength is long. The power of phasor technique again!

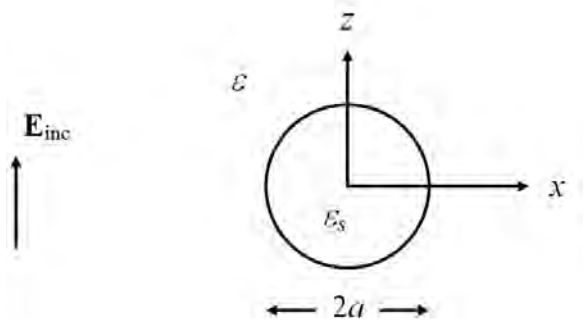


Figure 34.2: Geometry for studying the Rayleigh scattering problem.

34.1.1 Scattering by a Small Spherical Particle

The incident field polarizes the small particle making it look like a small electric dipole. Since the incident field is time harmonic, the small electric dipole will oscillate and radiate like a Hertzian dipole in the far field. First, we will look at the solution in the vicinity of the scatterer, namely, in the near field. Then we will motivate the form of the solution in the far field of the scatterer. (Solving a boundary value problem by looking at the solutions in two different physical regimes, and then matching (or patching) the solutions together is known as asymptotic matching or matched asymptotic expansions, a lost art [48].)

A Hertzian dipole can be approximated by a small current source so that

$$\mathbf{J}(\mathbf{r}) = \hat{z} I l \delta(\mathbf{r}) \tag{34.1.1}$$

Without loss of generality, we have assumed the Hertzian dipole to be at the origin. In the above, we let the time-harmonic current $I = dq/dt = j\omega q$. Then

$$Il = j\omega ql = j\omega p \tag{34.1.2}$$

where the dipole moment $p = ql$. As we have seen before, the vector potential $\mathbf{A}$ due to a Hertzian dipole, after substituting (34.1.1), is

$$\begin{aligned}\mathbf{A}(\mathbf{r}) &= \frac{\mu}{4\pi} \iiint_V d\mathbf{r}' \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} e^{-j\beta|\mathbf{r}-\mathbf{r}'|} \\ &= \hat{z}\frac{\mu I l}{4\pi r} e^{-j\beta r} = \hat{z}\frac{j\omega\mu q l}{4\pi r} e^{-j\beta r}\end{aligned} \tag{34.1.3}$$

where we have made use of the sifting property of the delta function in (34.1.1) when it is substituted into the above integral. The above is the exact solution due to a Hertzian dipole. The only approximation is to assume the dipole with current distribution given by (34.1.1), which is a good approximation if the dipole is much smaller than the wavelength.

Near Field

The above gives the vector potential $\mathbf{A}$ due to a Hertzian dipole. Since the dipole is infinitesimally small, the above solution is both valid in the near field as well as in the far field. Since the dipole moment ql is induced by the incident field, we need to relate ql to the amplitude of the incident electric field. To this end, we convert the above vector potential field to the near electric field of a small dipole.

From prior knowledge, we know that the electric field is given by $\mathbf{E} = -j\omega\mathbf{A} - \nabla\Phi$. From dimensional analysis, the scalar potential term dominates over the vector potential term in the near field of the scatterer since $\partial/\partial x \gg 1/c\partial/\partial t$. Hence, we need to derive, for the corresponding scalar potential, the approximate solution.

The scalar potential $\Phi(\mathbf{r})$ is obtained from the Lorenz gauge (see (27.2.23)) that $\nabla \cdot \mathbf{A} = -j\omega\mu\varepsilon\Phi$. Therefore,

$$\Phi(\mathbf{r}) = \frac{-1}{j\omega\mu\varepsilon}\nabla \cdot \mathbf{A} = -\frac{Il}{j\omega\varepsilon 4\pi}\frac{\partial}{\partial z}\frac{1}{r}e^{-j\beta r} \tag{34.1.4}$$

When we are close to the dipole, by assuming that $\beta r \ll 1$, we use a quasi-static approximation about the potential.[2] Then

$$\frac{\partial}{\partial z}\frac{1}{r}e^{-j\beta r} \approx \frac{\partial}{\partial z}\frac{1}{r} = \frac{\partial r}{\partial z}\frac{\partial}{\partial r}\frac{1}{r} = -\frac{z}{r}\frac{1}{r^2} \tag{34.1.5}$$

or after using that $z/r = \cos\theta$,

$$\Phi(\mathbf{r}) \approx \frac{ql}{4\pi\varepsilon r^2}\cos\theta \tag{34.1.6}$$

which is the static dipole potential because we are in the near field of the dipole. This dipole induced in the small particle is formed in response to the incident field and its dipole potential given by the previous expression. In other words, the incident field polarizes the small particle into a small time-oscillating Hertzian dipole which re-radiates.

Next, we calculate the polarizability of a small particle. The polarizability is a measure of the strength of the dipole moment induced by the incident field on a small particle. To find the polarizability, we need only to imagine the particle to be in between two parallel plates which are separated far apart, and with a constant electric field pointing in the z direction. Hence, the inducing (or incident) field can be approximated by a constant local static electric field,

$$\mathbf{E}_{inc} = \hat{z}E_i \tag{34.1.7}$$

This is the field that will polarize the small particle. It can also be an electric field between two parallel plates. The corresponding electrostatic potential for the inducing field is then

$$\Phi_{inc} = -zE_i \tag{34.1.8}$$

so that $\mathbf{E}_{inc} \approx -\nabla\Phi_{inc} = \hat{z}E_i$, as $\omega \to 0$. The scattered dipole potential from the spherical particle in the vicinity of it is quasi-static and is given by

$$\Phi_{sca} = E_s\frac{a^3}{r^2}\cos\theta \tag{34.1.9}$$

[2]This is the same as ignoring retardation effect.

which is the potential due to a static dipole. The electrostatic boundary value problem (BVP) has been previously solved and[3]

$$E_s = \frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon} E_i \tag{34.1.10}$$

Using (34.1.10) in (34.1.9), we get

$$\Phi_{sca} = \frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon} E_i \frac{a^3}{r^2} \cos\theta \tag{34.1.11}$$

On comparing with (34.1.6), one can see that the dipole moment induced by the incident field is given by

$$p = ql = 4\pi\varepsilon \frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon} a^3 E_i = \alpha E_i \tag{34.1.12}$$

where α is the polarizability of the small particle. The above analysis is valid as long as the particle size is much smaller than the wavelength. Hence, the incident field can be a time-harmonic field as well, even in the optical regime.

Far Field

Now, we have learnt that a small particle is polarized by the time-harmonic incident field, If the incident field is time-harmonic, the the small dipole will be time-oscillating and it will radiate like a time-varying Hertzian dipole whose far field we have previously derived (see Section 28.2). In the far field of the Hertzian dipole, we have

$$\mathbf{E} = -j\omega\mathbf{A} - \nabla\Phi = -j\omega\mathbf{A} - \frac{1}{j\omega\mu\varepsilon}\nabla\nabla\cdot\mathbf{A} \tag{34.1.13}$$

But when we are in the far field, $\mathbf{A}$ behaves like a spherical wave which in turn behaves like a local plane wave if one goes far enough. Therefore, $\nabla \to -j\boldsymbol{\beta} = -j\beta\hat{r}$. Using this approximation in (34.1.13), we arrive at

$$\mathbf{E} \cong -j\omega\left(\mathbf{A} - \frac{\boldsymbol{\beta}\boldsymbol{\beta}}{\beta^2}\cdot\mathbf{A}\right) = -j\omega(\mathbf{A} - \hat{r}\hat{r}\cdot\mathbf{A}) = -j\omega(\hat{\theta}A_\theta + \hat{\phi}A_\phi) \tag{34.1.14}$$

where we have used $\hat{r} = \boldsymbol{\beta}/\beta$. This is similar to the far field result we have derived in Section 29.1.2. In the above, we have used $\mathbf{A} = \overline{\mathbf{I}}\cdot\mathbf{A}$ and that $\overline{\mathbf{I}} = \hat{r}\hat{r} + \hat{\theta}\hat{\theta} + \hat{\phi}\phi$.

34.1.2 Scattering Cross Section

From (34.1.3), upon making use of (34.1.2), noticeably, $A_\phi = 0$ while

$$A_\theta = -\frac{j\omega\mu ql}{4\pi r} e^{-j\beta r}\sin\theta \tag{34.1.15}$$

[3]It was one of the homework problems. See also Section 8.3.7.

Consequently, using (34.1.12) for ql, we have in the far field that[4]

$$E_\theta \cong -j\omega A_\theta = -\frac{\omega^2\mu ql}{4\pi r}e^{-j\beta r}\sin\theta = -\omega^2\mu\varepsilon\left(\frac{\varepsilon_s-\varepsilon}{\varepsilon_s+2\varepsilon}\right)\frac{a^3}{r}E_i e^{-j\beta r}\sin\theta \tag{34.1.16}$$

Using local plane-wave approximation that

$$H_\phi \cong \sqrt{\frac{\varepsilon}{\mu}}E_\theta = \frac{1}{\eta}E_\theta \tag{34.1.17}$$

where $\eta = \sqrt{\mu/\varepsilon}$, the time-averaged Poynting vector is given by $\langle\mathbf{S}\rangle = 1/2\Re e\left\{\mathbf{E}\times\mathbf{H}^*\right\} = \frac{1}{2\eta}|E_\theta|^2\hat{r}$. Therefore, the total scattered power is obtained by integrating the power density over a spherical surface when r tends to infity. Thus, the total scattered power is

$$P_s = \int d\Omega\langle S_r\rangle = \frac{1}{2\eta}\int_0^\pi r^2\sin\theta d\theta\int_0^{2\pi}d\phi|E_\theta|^2 \tag{34.1.18}$$

$$= \frac{1}{2\eta}\beta^4\left|\frac{\varepsilon_s-\varepsilon}{\varepsilon_s+2\varepsilon}\right|^2\frac{a^6}{r^2}|E_i|^2r^2\left(\int_0^\pi\sin^3\theta d\theta\right)2\pi \tag{34.1.19}$$

But

$$\int_0^\pi\sin^3\theta d\theta = -\int_0^\pi\sin^2\theta d\cos\theta = -\int_0^\pi(1-\cos^2\theta)d\cos\theta$$
$$= -\int_1^{-1}(1-x^2)dx = \frac{4}{3} \tag{34.1.20}$$

Therefore

$$P_s = \frac{4\pi}{3\eta}\left|\frac{\varepsilon_s-\varepsilon}{\varepsilon_s+2\varepsilon}\right|^2\beta^4a^6|E_i|^2 \tag{34.1.21}$$

In the above, even though we have derived the equation using electrostatic theory, it is also valid for complex permittivity defined in Section 7.1.2. One can take the divergence of (7.1.11) to arrive at a Gauss' law for lossy dispersive media, viz., $\nabla\cdot\underline{\underline{\varepsilon}}\mathbf{E} = 0$ which is "homomorphic" to the lossless case. (Hurrah again to phasor technique!)

The scattering cross section is the effective area of a scatterer such that the total scattered power is proportional to the incident power density times the scattering cross section. As such it is defined as

$$\Sigma_s = \frac{P_s}{\langle S_{\rm inc}\rangle} = \frac{8\pi a^2}{3}\left|\frac{\varepsilon_s-\varepsilon}{\varepsilon_s+2\varepsilon}\right|^2(\beta a)^4 \tag{34.1.22}$$

where we have used the local plane-wave approximation that

$$\langle S_{\rm inc}\rangle = \frac{1}{2\eta}|E_i|^2 \tag{34.1.23}$$

[4]The ω^2 dependence in (34.1.16) of the far field implies that the radiated electric field in the far zone is proportional to the acceleration of the charges on the dipole.

The above also implies that

$$P_s = \langle S_{\text{inc}} \rangle \cdot \Sigma_s \tag{34.1.24}$$

In other words, the scattering cross section Σ_s is an effective cross-sectional area of the scatterer that will intercept the incident wave power $\langle S_{\text{inc}} \rangle$ to produce the scattered power P_s. This concept is similar to the effective aperture A_e of an antenna. In an antenna, we have

$$P_{\text{received}} = \langle S_{\text{inc}} \rangle \cdot A_e \tag{34.1.25}$$

where A_e is a measure of how much power is received or absorbed by the antenna. A similar idea of absorption cross-section area is also used in the community.

It is seen that the scattering cross section grows as the fourth power of frequency since $\beta = \omega/c$. The radiated field grows as the second power because it is proportional to the acceleration of the charges on the particle. The higher the frequency, the more the scattered power. This mechanism can be used to explain why the sky is blue. It also can be used to explain why sunset has a brilliant hue of red and orange (see Figure 34.3).

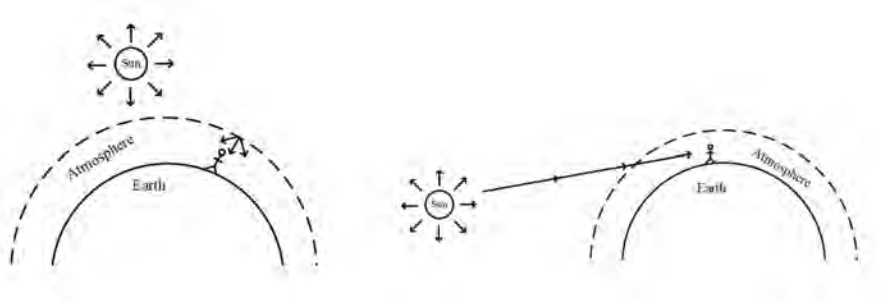


Figure 34.3: (Left) During the day time, when we look at the sky, we mainly see scattered sunlight. Since high-frequency light is scattered more, the sky appears blue. (Right) At sunset, the sunlight has to go through a thicker atmosphere. Thus, the blue light is scattered away, leaving the red light that reaches our eyes. Therefore, the color of the sunset appears red. (The figures are not drawn to scale.)

The above also explains the brilliant glitter of gold plasmonic nano-particles as discovered by ancient Roman artisans. For gold, the medium resembles a plasma, and hence, we can have $\varepsilon_s < 0$, and the denominator of (34.1.22) can be very small giving rise to strongly scattered light (see Section 8.3.7).

Furthermore, since the far field scattered power density of this particle is

$$\langle S \rangle = \frac{1}{2\eta} E_\theta H_\phi^* \sim \sin^2 \theta \tag{34.1.26}$$

the scattering power pattern of this small particle is not isotropic. In other words, these dipoles radiate predominantly in the broadside direction but not in their end-fire directions. Therefore, insects and sailors can use this to figure out where the sun is even in a cloudy day. In fact, it is like

a rainbow: If the sun is rising or setting in the horizon, there will be a bow across the sky where the scattered field is predominantly linearly polarized.[5] It is believed that the Vikings used such a "sunstone" for direction finding to traverse the Atlantic Ocean. A sunstone is shown in Figure 34.4.

Figure 34.4: A sunstone can indicate the polarization of the scattered light. From that, one can deduce where the sun is located; either behind us or in front of us (courtesy of Wikipedia).

34.1.3 Small Conductive Particle

The above analysis is for a small dielectric particle. The quasi-static analysis may not be valid for when the conductivity of the particle becomes very large. For instance, for a perfect electric conductor immersed in a time varying electromagnetic field, the magnetic field in the long wavelength limit induces eddy current in a PEC sphere.[6] Hence, in addition to an electric dipole component, a PEC sphere also has a magnetic dipole component. The scattered field due to a tiny PEC sphere is a linear superposition of an electric and magnetic dipole components. These two dipolar components have electric fields that cancel precisely at certain observation angle. It gives rise to deep null in the bi-static radar scattering cross-section (RCS)[7] of a PEC sphere as illustrated in Figure 34.5.

[5] You can go through a Gedanken experiment to convince yourself of such.

[6] Note that there is no PEC at optical frequencies. A metal behaves more like a plasma medium at optical frequencies.

[7] Scattering cross section in microwave range is called an RCS due to its prevalent use in radar technology.

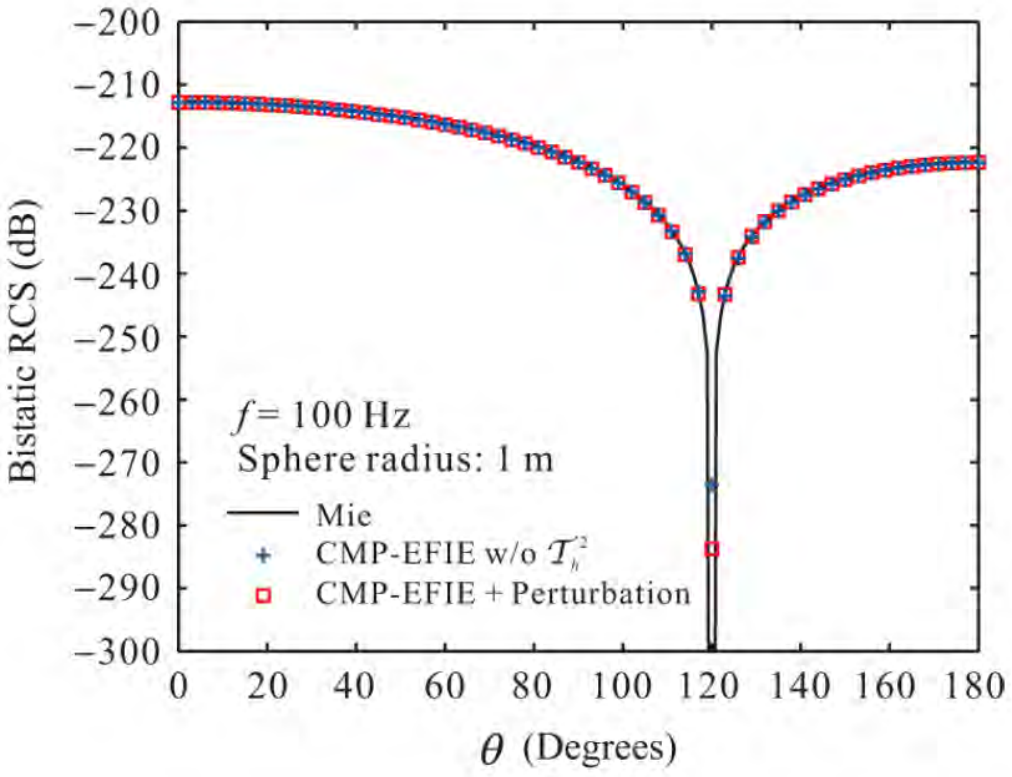


Figure 34.5: RCS (radar scattering cross section) of a small PEC scatterer (courtesy of Sheng et al. [254]).

34.2 Mie Scattering

When the size of the scatterer or the sphere becomes larger compared to wavelength λ, quasi-static approximation is insufficient to approximate the solution. Then one has to solve the boundary value problem in its full glory usually called the full-wave theory or Mie theory [255, 256]. With this theory, the scattering cross section does not grow indefinitely with frequency as in (34.1.22). It has to saturate to a value for increasing frequency. For a sphere of radius a, the scattering cross section becomes πa^2 in the high-frequency limit. This physical feature of this plot is shown in Figure 34.6, and it also explains why the sky is not purple.

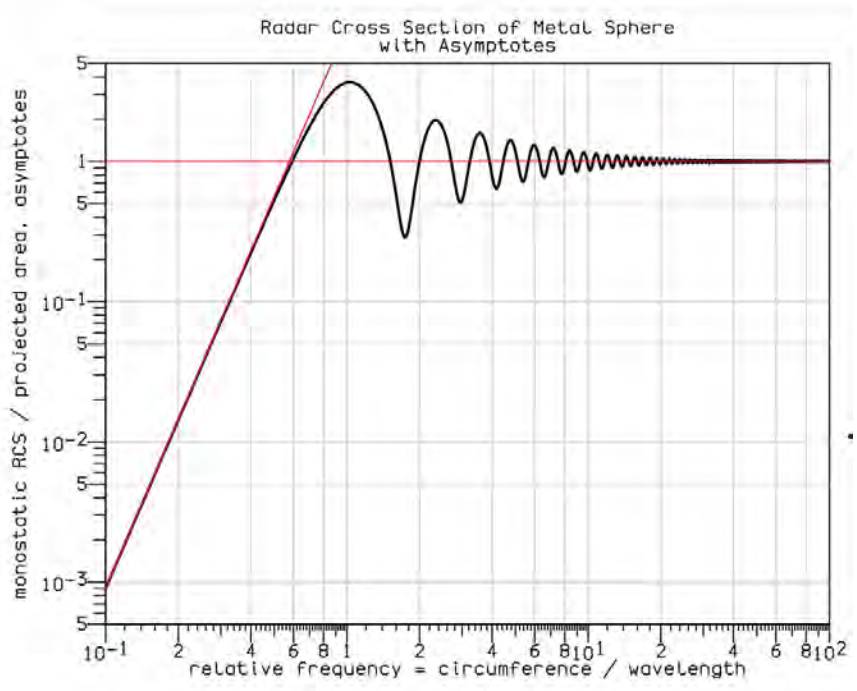


Figure 34.6: Radar cross section (RCS) calculated using Mie scattering theory [256].

34.2.1 Optical Theorem

Before we discuss the Mie scattering solution, let us discuss an amazing theorem called the optical theorem. This theorem says that the scattering cross section of a scatterer depends only on the forward scattering power density of the scatterer. In other words, if a plane wave is incident on a scatterer, the scatterer will scatter the incident wave and power in all directions. But the total power scattered by the object is only dependent on the forward scattering power density of the object or scatterer. This amazing theorem is called the optical theorem, and a proof of this is given in J.D. Jackson's book [49].

The true physical reason for this is power orthogonality. Two plane waves cannot interact or exchange power with each other unless they are propagating in the same direction, or they share the same $\mathbf{k}$ or $\boldsymbol{\beta}$ vector. When $\boldsymbol{\beta}$ is both the plane wave direction of the incident wave as well as the forward scattered wave, then power can be transferred from the incident plane wave to the forward scattered wave. This is similar to power orthogonality in a waveguide, where orthogonal modes of a waveguide carry power independently of each other [101, 90].

The scattering pattern of a scatterer for increasing frequency is shown in Figure 34.7. For Rayleigh scattering where the wavelength is long, the scattered power is distributed isotropically except for the doughnut shape of the radiation pattern, namely, the $\sin^2\theta$ dependence. As the frequency increases, the power is scattered increasingly in the forward direction. The reason being

that for very short wavelength, the scatterer looks like a disc to the incident wave, casting a shadow in the forward direction. Hence, there has to be scattered field in the forward direction to cancel the incident wave to cast this shadow. This point is often counter-intuitive to students of electromagnetics.

In a nutshell, the optical theorem is intuitively obvious for high-frequency scattering. The amazing part about this theorem is that it is true for all frequencies. But this is a consequence of power orthogonality, as mentioned above.

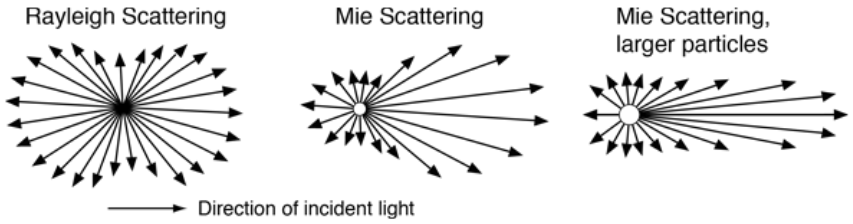


Figure 34.7: A particle scatters increasingly more in the forward direction as the frequency increases (Courtesy of hyperphysics.phy-astr.gsu.edu).

34.2.2 Mie Scattering by Spherical Harmonic Expansions

As explained above, as the wavelength becomes shorter, we need to solve the boundary value problem in its full glory without making any approximations like the previous section. This closed form solution can be found for a sphere scattering by using separation of variables and spherical harmonic expansions that will be discussed in the section.

The Mie scattering solution by a sphere will be discussed later in this chapter.[8] The separation of variables in spherical coordinates is not the only useful for Mie scattering, it is also useful for analyzing spherical cavity problems. So we will present the precursor knowledge so that you can read further into Mie scattering theory next.

34.2.3 Separation of Variables in Spherical Coordinates

To this end, we look at the scalar wave equation $(\nabla^2 + \beta^2)\Psi(\mathbf{r}) = 0$ in spherical coordinates. A lookup table can be used to evaluate $\nabla \cdot \nabla$, or divergence of a gradient in spherical coordinates. The Helmholtz wave equation then becomes[9]

$$\left(\frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial}{\partial r} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta} + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2} + \beta^2\right)\Psi(\mathbf{r}) = 0 \tag{34.2.1}$$

[8] It is also treated in J.A. Kong's book [34] and Chapter 3 of W.C. Chew, Waves and Fields in Inhomogeneous Media [37] and many other textbooks [49, 68, 104, 257].

[9] By quirk of mathematics, it turns out that the first term on the right-hand side below can be simplified by observing that $\frac{1}{r^2}\frac{\partial}{\partial r}r^2 = \frac{1}{r}\frac{\partial}{\partial r}r$.

Noting the $\partial^2/\partial\phi^2$ derivative, by using separation of variables technique, we assume $\Psi(\mathbf{r})$ to be of the form

$$\Psi(\mathbf{r}) = F(r,\theta)e^{jm\phi} \tag{34.2.2}$$

This will simplify the $\partial/\partial\phi$ derivative in the partial differential equation since $\frac{\partial^2}{\partial\phi^2}e^{jm\phi} = -m^2 e^{jm\phi}$. Then (34.2.1) becomes

$$\left(\frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial}{\partial r} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta} - \frac{m^2}{r^2\sin^2\theta} + \beta^2\right)F(r,\theta) = 0 \tag{34.2.3}$$

The above can be solved using the separation of variables, by letting that

$$F(r,\theta) = b_n(\beta r)P_n^m(\cos\theta) \tag{34.2.4}$$

where $b_n(\beta r)$ is a spherical Bessel function, and $P_n^m(\cos\theta)$ is the associate Legendre polynomial or function.[10] The associate Legendre's function satisfies

$$\left\{\frac{1}{\sin\theta}\frac{d}{d\theta}\sin\theta\frac{d}{d\theta} + \left[n(n+1) - \frac{m^2}{\sin^2\theta}\right]\right\}P_n^m(\cos\theta) = 0 \tag{34.2.5}$$

Note that (34.2.5) is an eigenvalue problem with eigenvalue $n(n+1)$, and $|m| \leq |n|$. The value $n(n+1)$ is also known as the separation constant.

Consequently, using (34.2.3), and (34.2.5), we can show that $b_n(\beta r)$ in (34.2.4) satisfies

$$\left[\frac{1}{r^2}\frac{d}{dr}r^2\frac{d}{dr} - \frac{n(n+1)}{r^2} + \beta^2\right]b_n(\beta r) = 0 \tag{34.2.6}$$

The above is the spherical Bessel equation where $b_n(\beta r)$ is either the spherical Bessel function $j_n(\beta r)$, spherical Neumann function $n_n(\beta r)$, or the spherical Hankel functions, $h_n^{(1)}(\beta r)$ and $h_n^{(2)}(\beta r)$. The spherical functions are the close cousins of the cylindrical functions $J_n(x)$, $N_n(x)$, $H_n^{(1)}(x)$ and $H_n^{(2)}(x)$. They are related to the cylindrical functions via [52, 37] [11]

$$b_n(\beta r) = \sqrt{\frac{\pi}{2\beta r}}B_{n+\frac{1}{2}}(\beta r) \tag{34.2.7}$$

Since (34.2.6) is a second order ordinary differential equation, only two of the four possible solutions are independent similar to the cylindrical Bessel functions case [108, p. 14], only two of the possible four solutions are linearly independent.

It is customary to define the spherical harmonics as [49, 108]

$$Y_{nm}(\theta,\phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}P_n^m(\cos\theta)e^{jm\phi} \tag{34.2.8}$$

[10] Thanks goodness, these special functions like Legendre polynomial and Bessel functions were discovered in the late 1700s before Maxwell's equations were discovered!

[11] By a quirk of nature, the spherical Bessel functions needed for 3D wave equations are in fact simpler than cylindrical Bessel functions needed for 2D wave equation. One can say that 3D is real, but 2D is surreal.

The above is ortho-normalized such that

$$Y_{n,-m}(\theta,\phi) = (-1)^m Y^*_{nm}(\theta,\phi) \tag{34.2.9}$$

and that

$$\int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta d\theta Y^*_{n'm'}(\theta,\phi) Y_{nm}(\theta,\phi) = \delta_{n'n}\delta_{m'm} \tag{34.2.10}$$

These functions are also complete[12] like Fourier series, so that

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n} Y^*_{nm}(\theta',\phi') Y_{nm}(\theta,\phi) = \delta(\phi-\phi^{'})\delta(\cos\theta - \cos\theta^{'}) \tag{34.2.11}$$

In general, the solution to the scalar wave equation in spherical coordinates can be expanded into the form

$$\Psi(\mathbf{r}) = \Psi(r,\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} [A_{nm} j_n(kr) + B_{nm} n_n(kr)] Y_{nm}(\theta,\phi) \tag{34.2.12}$$

If the solution space includes the origin, we set $B_{nm} = 0$ since $n_n(kr)$ is singular at $r = 0$, unless a source exists at $r = 0$. In general, the orthonormality of the spherical harmonics expressed in (34.2.11) can be used to find the expansion coefficients A_{nm} and B_{nm}.

[12] In a nutshell, a set of basis functions is complete in a subspace if any function in the same subspace can be expanded as a sum of these basis functions.

34.3 More on Mie Scattering

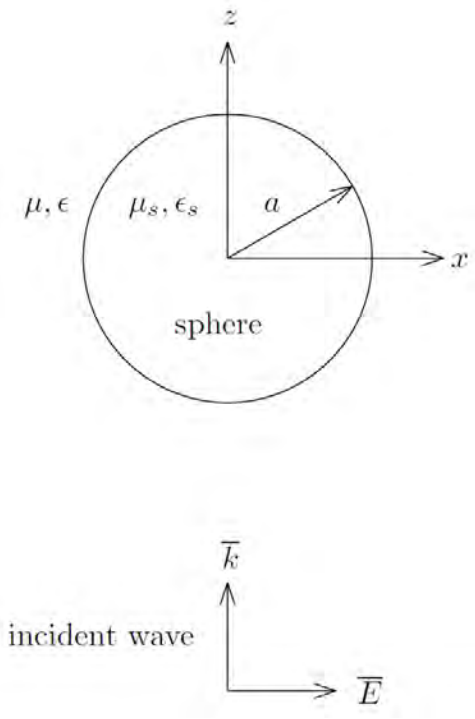


Figure 34.8: Geometry to illustrate the Mie scattering problem. The incident wave is traveling in the z direction so that it is symmetric about the z axis which simplifies the analysis. The problem can be solved in closed form using spherical harmonic expansions and Debye potentials.

The solution of electromagnetic scattering by a spherical scatterer can be found exactly. To do this, we need to define the Debye potential π_e and π_m [255]. These are like the pilot potential we have defined earlier for hollow waveguides. Unlike a hollow waveguide, the pilot vector is not a constant vector. In spherical coordinates, to characterize the TM to r and TE to r waves, we use pilot vector $\mathbf{r}$. Thus, for TM to r waves, we have

$$\mathbf{H}^{TM} = \nabla \times \mathbf{r}\pi_e \tag{34.3.1}$$

and for TE to r waves, we have

$$\mathbf{E}^{TE} = \nabla \times \mathbf{r}\pi_m. \tag{34.3.2}$$

Using the fact that the above fields are solutions to Helmholtz wave equations, it can be shown that [108]

$$\left(\nabla^2 + k^2\right)\pi_e = 0, \tag{34.3.3}$$

$$(\nabla^2 + k^2)\,\pi_m = 0. \tag{34.3.4}$$

The above is incredibly simple even though we have started with a non-constant pilot vector $\mathbf{r}$.

In general, in a source-free region, we can derive the electromagnetic fields as

$$\mathbf{H} = \nabla \times \mathbf{r}\pi_e + \frac{1}{-j\omega\mu}\nabla \times \nabla \times \mathbf{r}\pi_m, \tag{34.3.5}$$

$$\mathbf{E} = \nabla \times \mathbf{r}\pi_m + \frac{1}{j\omega\epsilon}\nabla \times \nabla \times \mathbf{r}\pi_e. \tag{34.3.6}$$

Furthermore, we can extract the r components of $\mathbf{H}$ and $\mathbf{E}$ in the above to yield [258, Exercise 3.22, Solution Manual]

$$H_r = -\frac{1}{j\omega\mu}\left[\frac{\partial^2}{\partial r^2} r\pi_m + k^2 r\pi_m\right], \tag{34.3.7}$$

$$E_r = \frac{1}{j\omega\epsilon}\left[\frac{\partial^2}{\partial r^2} r\pi_e + k^2 r\pi_e\right]. \tag{34.3.8}$$

Since π_m and π_e are solutions of the scalar wave equations (34.3.3) and (34.3.4), their general solutions are of the form

$$\begin{Bmatrix} j_n(kr) \\ h_n^{(2)}(kr) \end{Bmatrix} P_n^m(\cos\theta) \begin{Bmatrix} \cos m\phi \\ \sin m\phi \end{Bmatrix} \tag{34.3.9}$$

where the braces imply "linear superpositions of." In the above, $j_n(x)$ is the spherical Bessel function and $h_n^{(2)}(x)$ is the spherical Hankel function of the second kind. The above is the spherical harmonics representation of the wave solution in spherical coordinates.

Since π_m and π_e are of the form given in (34.3.9), then from the scalar wave equation expressed in spherical coordinates, we can show that (34.3.7) and (34.3.8) simplify to

$$H_r = \frac{1}{-j\omega\mu}\frac{n(n+1)}{r}\pi_m \tag{34.3.10}$$

$$E_r = \frac{1}{j\omega\varepsilon}\frac{n(n+1)}{r}\pi_e \tag{34.3.11}$$

In other words, H_r and E_r uniquely define π_m and π_e

Interface Boundary Conditions

The interface boundary conditions for a spherical surface are that tangential E and H are continuous.

By extracting the transverse to r components of (34.3.5) and (34.3.6), we have

$$\mathbf{H}_s = -\mathbf{r} \times \nabla_s \pi_e + \frac{1}{i\omega\mu}\frac{1}{r}\frac{\partial}{\partial r} r^2 \nabla_s \pi_m, \tag{34.3.12}$$

$$\mathbf{E}_s = -\mathbf{r} \times \nabla_s \pi_m - \frac{1}{i\omega\epsilon}\frac{1}{r}\frac{\partial}{\partial r} r^2 \nabla_s \pi_e. \tag{34.3.13}$$

One can easily show that if we have only TE to r or TM to r waves, the boundary conditions can be satisfied still. This implies that the boundary conditions for continuous tangential E and tangential H fields do not couple the TE and TM waves. Or the TE wave solution can be solved for independently of the TM wave solution![13]

It can be shown that to ensure continuity of tangential E and H fields, it is necessary to have the continuity of

$$\pi_m, \quad \pi_e \tag{34.3.14}$$

and the continuity of

$$\frac{1}{r\varepsilon}\frac{\partial}{\partial r}r\pi_e, \quad \frac{1}{r\mu}\frac{\partial}{\partial r}r\pi_m \tag{34.3.15}$$

To re-emphasize, the TE and TM boundary conditions can be satisfied independently of each other. We will take advantage of this to simplify the solution to this problem and solve the TE and TM scattering independently of each other.

Therefore, to solve the sphere scattering problem of a plane wave, we need first to expand the plane wave in terms of spherical harmonics. Then we need to extract the TE to r or TM to r waves or the Debye potentials.

First, if we have a plane wave polarized in the $\hat{x}$ direction incident on the spherical scatterer, then

$$\mathbf{E} = \hat{x}E_0 e^{-jkz} = \hat{x}E_0 e^{-jkr\cos\theta} \tag{34.3.16}$$

$$\mathbf{H} = \hat{y}\sqrt{\frac{\varepsilon}{\mu}}E_0 e^{-jkr\cos\theta} \tag{34.3.17}$$

Further, we need to use the *wave transformation* that a plane wave can be expressed in terms of spherical harmonics. Namely,

$$e^{-jkr\cos\theta} = \sum_{n=0}^{\infty} j^{-n}(2n+1)j_n(kr)P_n(\cos\theta) \tag{34.3.18}$$

The incident wave is chosen so that it is axially symmetric simplifying the spherical harmonics expansion: only $m = 0$ harmonic is needed (see (34.2.12)). (The above is proved in Harrington [53].) Thus we can write the incident field in (34.3.16), after using $\hat{r}\cdot\hat{x} = \sin\theta\cos\phi$ as

$$E_r^{\text{inc}} = E_0 \sin\theta\cos\phi e^{-jkr\cos\theta} \tag{34.3.19}$$

$$= E_0 \frac{\cos\phi}{jkr}\frac{\partial}{\partial\theta}\left(e^{-jkr\cos\theta}\right) \tag{34.3.20}$$

Thus we expressed the incident electric field, after using the wave transformation, as

$$E_r^{\text{inc}} = \frac{E_0\cos\phi}{jkr}\sum_{n=0}^{\infty} j^{-n}(2n+1)j_n(kr)\frac{\partial}{\partial\theta}P_n(\cos\theta) \tag{34.3.21}$$

[13] It is a mystery of nature that wave problems in odd dimensions are simpler than problems in even dimensions!

Then using $\frac{\partial}{\partial\theta}P_n(\cos\phi) = P_n^1(\cos\theta)$, the above becomes

$$E_r^{\text{inc}} = -j\frac{E_0\cos\phi}{(kr)^2}\sum_{n=1}^{\infty} j^{-n}(2n+1)\hat{J}_n(kr)P_n^1(\cos\theta) \tag{34.3.22}$$

The summation begins with $n = 1$ because $P_0^1(\cos\theta) = 0$. Also, we define $\hat{J}_n(kr) = krj_n(kr)$ which is the normalized spherical Bessel function.

Next, we need to identify the TE and TM waves buried in the incident wave. Comparing the above equations (34.3.10) and (34.3.11), we identify the Debye potential for the TM and TE polarizations, given the $\mathbf{r}$ component of the fields. Thus for the incident fields, the Debye potentials are identified as

$$\pi_e^{\text{inc}} = -\frac{E_0\cos\phi}{\omega\mu r}\sum_{n=1}^{\infty} j^{-n}\frac{(2n+1)}{n(n+1)}\hat{J}_n(kr)P_n^1(\cos\theta) \tag{34.3.23}$$

$$\pi_m^{\text{inc}} = \frac{E_0\sin\phi}{kr}\sum_{n=1}^{\infty} j^{-n}\frac{(2n+1)}{n(n+1)}\hat{J}_n(kr)P_n^1(\cos\theta) \tag{34.3.24}$$

Next, we write down the Debye potentials for the TM scattered fields to be

$$\pi_e^s = \frac{E_0\cos\phi}{\omega\mu r}\sum_{n=1}^{\infty} a_n\hat{H}_n^{(2)}(kr)P_n^1(\cos\theta) \tag{34.3.25}$$

$$\pi_m^s = \frac{E_0\sin\phi}{kr}\sum_{n=1}^{\infty} b_n\hat{H}_n^{(2)}(kr)P_n^1(\cos\theta) \tag{34.3.26}$$

Here, $\hat{H}_n^{(2)}(x)$ are normalized spherical Hankel functions of the second kind. These scattered fields have to be outgoing waves when $kr \to \infty$. Hence, spherical Hankel functions of the second kind are chosen with $\exp(j\omega t)$ time dependence.

In addition, we need to write down the solutions inside the spherical scatterer. They have to be regular at the origin or at $r = 0$. Therefore, normalized spherical Bessel functions $\hat{J}_n(x)$ are chosen as the solutions. Consequently,

$$\pi_e^i = \frac{E_0\cos\phi}{\omega\mu_s r}\sum_{n=1}^{\infty} c_n\hat{J}_n(k_s r)P_n^1(\cos\theta) \tag{34.3.27}$$

$$\pi_m^i = \frac{E_0\sin\phi}{k_s r}\sum_{n=1}^{\infty} d_n\hat{J}_n(k_s r)P_n^1(\cos\theta) \tag{34.3.28}$$

As discussed before, the scattering by a spherical scatter can be solved as two independent problems of TE to r and TM to r waves. Thus, after matching boundary conditions that π_e and $\frac{1}{\varepsilon}\frac{\partial}{\partial r}r\pi_e$ (and their dual for the TM case) are continuous at the interface,[14] we can show that

$$a_n = \frac{j^{-n}(2n+1)}{n(n+1)}\frac{-\sqrt{\varepsilon_s\mu}\hat{J}_n'(ka)\hat{J}_n(k_s a) + \sqrt{\varepsilon\mu_s}\hat{J}_n(ka)\hat{J}_n'(k_s a)}{\sqrt{\varepsilon_s\mu}\hat{H}_n^{(2)'}(ka)\hat{J}_n(k_s a) - \sqrt{\varepsilon\mu_s}\hat{H}_n^{(2)}(ka)\hat{J}_n'(k_s a)} \tag{34.3.29}$$

[14]These boundary conditions also follow from eqs. (3.8.1) of [108].

$$b_n = \frac{j^{-n}(2n+1)}{n(n+1)} \frac{-\sqrt{\varepsilon_s\mu}\hat{J}_n(ka)\hat{J}'_n(k_sa) + \sqrt{\varepsilon\mu_s}\hat{J}'_n(ka)\hat{J}_n(k_sa)}{\sqrt{\varepsilon_s\mu}\hat{H}^{(2)}_n(ka)\hat{J}'_n(k_sa) - \sqrt{\varepsilon\mu_s}\hat{H}^{(2)'}_n(ka)\hat{J}_n(k_sa)} \tag{34.3.30}$$

$$c_n = \frac{j^{-n}(2n+1)}{n(n+1)} \frac{-j\sqrt{\varepsilon_s\mu}}{\sqrt{\varepsilon_s\mu}\hat{H}^{(2)'}_n(ka)\hat{J}_n(k_sa) - \sqrt{\varepsilon\mu_s}\hat{H}^{(2)}_n(ka)\hat{J}'_n(k_sa)} \tag{34.3.31}$$

$$d_n = \frac{j^{-n}(2n+1)}{n(n+1)} \frac{j\sqrt{\varepsilon_s\mu}}{\sqrt{\varepsilon_s\mu}\hat{H}^{(2)}_n(ka)\hat{J}'_n(k_sa) - \sqrt{\varepsilon\mu_s}\hat{H}^{(2)'}_n(ka)\hat{J}_n(k_sa)} \tag{34.3.32}$$

The Wronskian of spherical Bessel functions has been used to simplify the above [49], [37, p. 189], viz.,

$$\hat{J}_n(k_1a)\hat{H}^{(2)'}_n(ka) - \hat{J}'_n(ka)\hat{H}^{(2)}_n(ka) = -j$$

These solutions, complicated though they are, are very useful because they are the exact solution of a scattering by a spherical object in electromagnetics. They are useful for validating numerical solutions of scattering problems.

Exercises for Lecture 34

Problem 34-1:

(i) Explain how the Vikings could have used the physical results of Rayleigh scattering to navigate themselves across the North Atlantic Ocean to arrive at Iceland.

(ii) Verify (34.1.22), and explain why it is also valid for complex permittivity case.

(iii) Give an intuitive reason as to why the optical theorem makes sense in the high-frequency limit.

(iv) By using the separation of variables, explain how you would solve the Helmholtz wave equation in 3D in spherical coordinates.

$$\left(\nabla^2 + \beta^2\right) \Psi(\mathbf{r}) = 0$$

(v) By using the interface boundary conditions suggested before (34.3.29), derive the constants a_n, b_n, c_n, and d_n and verify if they are correct.

Chapter 35

Spectral Expansions of Source Fields—Sommerfeld Integrals

In previous lectures, we have assumed plane waves in finding closed form solutions. Plane waves are simple waves, and their closed form solutions for reflections off a flat surface or a planarly layered medium can be found easily. But plane waves are figments of mathematical imaginations that are not encountered in the real world.

All sources can be thought of as linear superposition of point sources. Thus it is important to study the point source solution first. When the source is a point source, it generates a spherical wave. We do not know how to reflect a spherical wave off a planar interface exactly, but we do know how to reflect plane waves off a planar interfaces. But by expanding a spherical wave as a sum of plane waves and evanescent waves using Fourier transform technique, we can solve for the solution of a point source over a layered medium easily in terms of spectral integrals using Fourier transform in space. But these integrals are complicated; and Sommerfeld was the first person to have done this, and hence, these integrals are often called Sommerfeld integrals.

Finally, we shall apply the method of stationary phase to approximate these complicated integrals to elucidate their physics. From this, we can see ray physics and Fermat's principle emerging from the complicated mathematics. It reminds us of a lyric from the musical The Sound of Music—Ray, a drop of golden sun! Ray has mesmerized the human mind, and it will be interesting to see if the mathematics behind it is equally enchanting.

By this time, you probably feel inundated by the ocean of knowledge that you are imbibing. But if you can assimilate them, it will be an exhilarating experience as the knowledge will last a lifetime. Also, in the latter part of the course, we will use $e^{-i\omega t}$ time convention as is often used by mathematicians, physicists, and the optics community.

35.1 Spectral Representations of Sources

As mentioned above, a plane wave is a mathematical idealization that does not exist in the real world. In practice, waves are nonplanar in nature as they are generated by finite sources, such as antennas and scatterers: For example, a point source generates a spherical wave which is nonplanar.

Fortunately, these non-planar waves can be expanded as sum of plane waves. Once this is done, then the study of non-plane-wave reflections from a layered medium becomes routine.

In the following, we shall show how waves resulting from a point source can be expanded in terms of a plane wave summation (or integral). This topic is found in many textbooks [248, 139, 259, 1, 140, 37, 34, 104].

35.1.1 A Point Source—Fourier Expansion and Contour Integration

There are a number of ways to derive the plane wave expansion of a point source [108][Chap. 2]. We will illustrate one of the ways. The Fourier expansion in space, or spectral decomposition, or the plane-wave expansion of the field due to a point source could be derived using Fourier transform technique. First, notice that the scalar wave equation for the field due to a point source at the origin is

$$\left(\nabla^2 + k_0^2\right)\phi(x,y,z) = \left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} + k_0^2\right]\phi(x,y,z) = -\delta(x)\,\delta(y)\,\delta(z). \tag{35.1.1}$$

The above equation could be solved for the field $\phi(x,y,z)$ in the spherical coordinates, yielding the solution given in the previous lecture, namely, Green's function with the source point at the origin, or that[1]

$$\phi(x,y,z) = \phi(r) = \frac{e^{ik_0 r}}{4\pi r}. \tag{35.1.2}$$

The solution is entirely spherically symmetric due to the symmetry and location of the point source.

Next, assuming that the Fourier transform of $\phi(x,y,z)$ exists,[2] we can write

$$\phi(x,y,z) = \frac{1}{(2\pi)^3}\iiint_{-\infty}^{\infty} dk_x dk_y dk_z\, \tilde{\phi}(k_x,k_y,k_z) e^{ik_x x + ik_y y + ik_z z}. \tag{35.1.3}$$

Then we substitute the above into the left-hand side of (35.1.1), after exchanging the order of differentiation and integration,[3] one can simplify the Laplacian operator in the Fourier space, or spectral domain, to arrive at

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \overbrace{\longrightarrow}^{\text{Fourier Transform}} -k_x^2 - k_y^2 - k_z^2$$

Then, together with the Fourier representation of the delta function, on the right-hand side of (35.1.1), which is[4]

$$\delta(x)\,\delta(y)\,\delta(z) = \frac{1}{(2\pi)^3}\iiint_{-\infty}^{\infty} dk_x dk_y dk_z\, e^{ik_x x + ik_y y + ik_z z} \tag{35.1.4}$$

[1] From this point onward, we will adopt the $\exp(-i\omega t)$ time convention to be commensurate with the optics and physics literatures.

[2] The Fourier transform of a function $f(x)$ exists if it is absolutely integrable, namely that $\int_{-\infty}^{\infty} |f(x)|dx$ is finite (see [108]).

[3] Exchanging the order of differentiation and integration is allowed if the integral converges after the exchange.

[4] We have made use of that $\delta(x) = 1/(2\pi)\int_{-\infty}^{\infty} dk_x \exp(ik_x x)$ three times.

we convert (35.1.1) into

$$\iiint_{-\infty}^{\infty} dk_x dk_y dk_z \, [k_0^2 - k_x^2 - k_y^2 - k_z^2]\tilde{\phi}(k_x, k_y, k_z) e^{ik_x x + ik_y y + ik_z z}$$

$$= - \iiint_{-\infty}^{\infty} dk_x dk_y dk_z \, e^{ik_x x + ik_y y + ik_z z}. \tag{35.1.5}$$

Since the above is equal for all x, y, and z, we deduce that their integrands are the same. Hence, we arrive at inverse transform the above to get

$$\tilde{\phi}(k_x, k_y, k_z) = \frac{-1}{k_0^2 - k_x^2 - k_y^2 - k_z^2}. \tag{35.1.6}$$

Consequently, using this in (35.1.3), we have that $\phi(x, y, z)$ expressed in terms of the Fourier inverse transform, viz.,

$$\phi(x, y, z) = \frac{-1}{(2\pi)^3} \iiint_{-\infty}^{\infty} d\mathbf{k} \, \frac{e^{ik_x x + ik_y y + ik_z z}}{k_0^2 - k_x^2 - k_y^2 - k_z^2}. \tag{35.1.7}$$

where $d\mathbf{k} = dk_x dk_y dk_z$. The above expresses the fact the $\phi(x, y, z)$ which is a spherical wave by (35.1.2), is expressed as an integral summation of "plane waves". But these "plane waves" are not physical plane waves in free space since $k_x^2 + k_y^2 + k_z^2 \neq k_0^2$. In other words, these "plane waves" do not satisfy the physical dispersion relation of a plane wave.

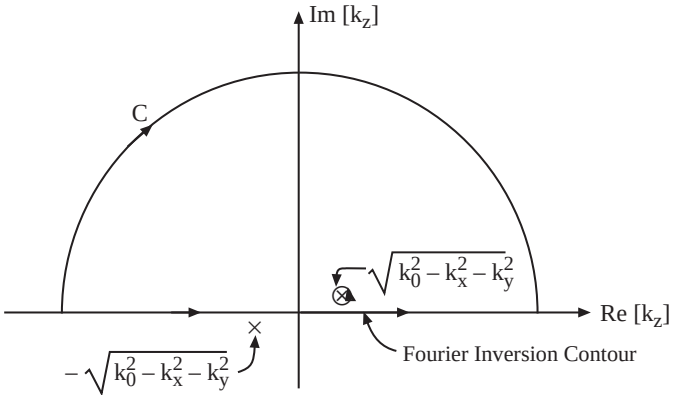


Figure 35.1: By invoking Cauchy's residue theorem, the integration along the real axis is equal to the integration along C plus the residue of the pole at $(k_0^2 - k_x^2 - k_y^2)^{1/2}$. By invoking Jordan's lemma, the integration around C is zero for most integrals, and thus, we require only the residue evaluation.

Weyl Identity—Plane-Wave Expansion of a Point-Source Field

To make the plane waves in (35.1.7) into physical plane waves, we have to massage it into a different form. We rearrange the triple integrals in (35.1.7) so that the dk_z integral is performed first. In other words,

$$\phi(\mathbf{r}) = \frac{1}{(2\pi)^3} \iint_{-\infty}^{\infty} dk_x dk_y e^{ik_x x + ik_y y} \int_{-\infty}^{\infty} dk_z \frac{e^{ik_z z}}{k_z^2 - (k_0^2 - k_x^2 - k_y^2)} \tag{35.1.8}$$

where we have deliberately rearrange the denominator with k_z being the variable in the inner integral. Then the integrand has poles at $k_z = \pm(k_0^2 - k_x^2 - k_y^2)^{1/2}$.[5] Moreover, for real k_0, and real values of k_x and k_y, these two poles lie on the real axis, rendering the integral in (35.1.7) undefined. However, if a small loss is assumed in k_0 such that $k_0 = k_0' + ik_0''$, then the poles migrate to be off the real axis (see Figure 35.1), and the integrals in (35.1.8) are well-defined. (In actual fact, this is intimately related to the uniqueness principle we have studied before: An infinitesimal loss is needed to guarantee uniqueness in an open space as shall be explained below.)

First, the reason is that without loss, $|\phi(\mathbf{r})| \sim O(1/r)$, $r \to \infty$ is not strictly absolutely integrable; and hence, its Fourier transform does not exist [55]: The manipulation that leads to (35.1.7) is not strictly correct. Second, the introduction of a small loss also guarantees the Sommerfeld radiation condition and the uniqueness of the solution to (35.1.1), and therefore, the equality of (35.1.2) and (35.1.8) [37].

Observe that in (35.1.8), when $z > 0$, the integrand is exponentially small when $\Im m[k_z] \to \infty$. Therefore, by Jordan's lemma [92], the integration for k_z vanishes over the contour C as shown in Figure 35.1. Then, by Cauchy's theorem [92], the integration over the Fourier inversion contour on the real axis is the same as integrating around the pole singularity located at $(k_0^2 - k_x^2 - k_y^2)^{1/2}$, yielding the residue of the pole (see Figure 35.1). Consequently, after doing the residue evaluation, a triple integral becomes a double integral, and we have

$$\phi(x, y, z) = \frac{i}{2(2\pi)^2} \iint_{-\infty}^{\infty} dk_x dk_y \frac{e^{ik_x x + ik_y y + ik_z' z}}{k_z'}, \quad z > 0, \tag{35.1.9}$$

where $k_z' = (k_0^2 - k_x^2 - k_y^2)^{1/2}$ is the value of k_z at the pole location.

Similarly, for $z < 0$, we can add a contour C in the lower-half plane that contributes zero to the integral. Therefore, one can deform the contour to pick up the pole contribution. As such, the integral is equal to the pole contribution at $k_z' = -(k_0^2 - k_x^2 - k_y^2)^{1/2}$ (see Figure 35.1). A similar expression similar to (35.1.9) can be derived for $z < 0$. As such, the result valid for all z can be written as

$$\phi(x, y, z) = \frac{i}{2(2\pi)^2} \iint_{-\infty}^{\infty} dk_x dk_y \frac{e^{ik_x x + ik_y y + ik_z' |z|}}{k_z'}, \quad \text{all } z. \tag{35.1.10}$$

[5] In (35.1.7), the pole is located at $k_x^2 + k_y^2 + k_z^2 = k_0^2$. This equation describes a sphere in $\mathbf{k}$ space, known as the Ewald's sphere [260].

By the uniqueness of the solution to the partial differential equation (35.1.1) satisfying radiation condition at infinity, we can equate (35.1.2) and (35.1.10), yielding the identity

$$\frac{e^{ik_0 r}}{r} = \frac{i}{2\pi} \iint_{-\infty}^{\infty} dk_x dk_y \, \frac{e^{ik_x x + ik_y y + ik_z |z|}}{k_z}, \tag{35.1.11}$$

where $k_x^2 + k_y^2 + k_z^2 = k_0^2$, or $k_z = (k_0^2 - k_x^2 - k_y^2)^{1/2}$. Now the plane wave in the integrand of the above is a physical plane wave. The above is known as the *Weyl identity* (Weyl 1919)[6]. To ensure the radiation condition, we require that $\Im m[k_z] > 0$ and $\Re e[k_z] > 0$ over all values of k_x and k_y in the integration. Furthermore, Equation (35.1.11) could be interpreted as an integral summation of physical plane waves propagating in all directions, including evanescent waves. It is the plane-wave expansion (including evanescent wave) of a spherical wave.

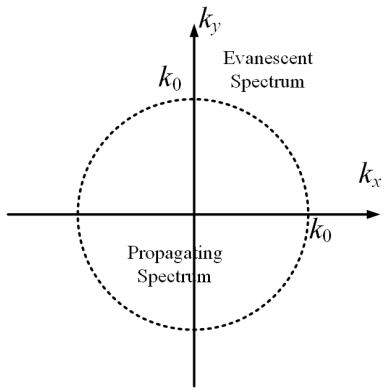


Figure 35.2: The integral in the Weyl identity is done over the entire k_x-k_y plane. The wave is propagating for $\mathbf{k}_\rho = \hat{x}k_x + \hat{y}k_y$ vectors inside the disk, while the wave is evanescent for $\mathbf{k}_\rho$ outside the disk. This is easily seen by the $\exp(ik_z|z|) = \exp(i\sqrt{k_0^2 - k_x^2 - k_y^2}|z|)$ dependence of the integrand in (35.1.11). Therefore, the high spectral frequency component of the Fourier integral is outside the disk.

One can also interpret the above as a 2D integral in the Fourier space over the k_x-k_y plane or variables. When $k_x^2 + k_y^2 < k_0^2$, or the spatial spectrum involving k_x and k_y is inside a disk of radius k_0, the waves are propagating waves. But for contributions outside this disk, the waves are evanescent (see Figure 35.2). And the high Fourier (or spectral) components of the Fourier spectrum correspond to evanescent waves. These high spectral components, which are related to the evanescent waves, are important for reconstituting the singularity of the Green's function.[7] Unfortunately, they become exponentially small the further we are from the source point.

[6]You will notice that this was derived after Sommerfeld had solved his problem, implying that Sommerfeld should have known this identity before Weyl.

[7]It may be difficult to wrap your head around so many new concepts, and you will have to contemplate on them to deeply understand them.

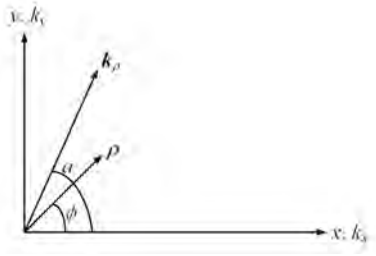


Figure 35.3: The $\mathbf{k}_\rho$ and the $\boldsymbol{\rho}$ vectors on the k_x-k_y plane and the x-y plane. The two planes are superposed in the same figure in this picture. The 2D plane wave is given by $\exp(i\mathbf{k}_\rho \cdot \boldsymbol{\rho})$.

Sommerfeld Identity—A Semi-Infinite Integral

The Weyl identity has double integral, and hence, is more difficult to integrate numerically. Here, we shall derive the Sommerfeld identity which has only one semi-infinite integral. First, in (35.1.11), we express the integral in cylindrical coordinates and write $\mathbf{k}_\rho = \hat{x}k_\rho \cos\alpha + \hat{y}k_\rho \sin\alpha$, $\boldsymbol{\rho} = \hat{x}\rho\cos\phi + \hat{y}\rho\sin\phi$ (see Figure 35.3). Then in cylindrical coordinates, $dk_x dk_y = k_\rho dk_\rho\, d\alpha$, and $k_x x + k_y y = \mathbf{k}_\rho \cdot \boldsymbol{\rho} = k_\rho \cos(\alpha - \phi)$, and with the appropriate change of variables, (35.1.11) becomes

$$\frac{e^{ik_0 r}}{r} = \frac{i}{2\pi} \int_0^\infty k_\rho dk_\rho \int_0^{2\pi} d\alpha \frac{e^{ik_\rho \rho\cos(\alpha-\phi)+ik_z|z|}}{k_z}, \tag{35.1.12}$$

where $k_z = (k_0^2 - k_x^2 - k_y^2)^{1/2} = (k_0^2 - k_\rho^2)^{1/2}$, where in cylindrical coordinates, in the $\mathbf{k}_\rho$-space, or the Fourier space, $k_\rho^2 = k_x^2 + k_y^2$. As such, using the integral identity for Bessel functions given by[8]

$$J_0(k_\rho\rho) = \frac{1}{2\pi}\int_0^{2\pi} d\alpha\, e^{ik_\rho\rho\cos(\alpha-\phi)}, \tag{35.1.13}$$

We can use the above to replace the $d\alpha$ integral in (35.1.12). Therefore, (35.1.12) becomes

$$\frac{e^{ik_0 r}}{r} = i \int_0^\infty dk_\rho \frac{k_\rho}{k_z} J_0(k_\rho\rho) e^{ik_z|z|}. \tag{35.1.14}$$

The above is also known as the *Sommerfeld identity* (Sommerfeld 1909 [155]; [248, p. 242]). Its physical interpretation is that a spherical wave can now be expanded as an integral summation of conical waves or cylindrical waves in the ρ direction, times a plane wave in the z direction over all wave numbers k_ρ. This wave is propagating in the $\pm z$ direction when $k_\rho < k_0$, but evanescent in the $\pm z$ direction when $k_\rho > k_0$ as shown in Figure 35.2.

[8]See Chew [37][eq. (2.2.15)], or Whitaker and Watson(1927) [261].

By using the fact that $J_0(k_\rho\rho) = 1/2[H_0^{(1)}(k_\rho\rho) + H_0^{(2)}(k_\rho\rho)]$, and the reflection formula that $H_0^{(1)}(e^{i\pi}x) = -H_0^{(2)}(x)$,[9] Then a variation of the above identity can be derived as [37]

$$\frac{e^{ik_0r}}{r} = \frac{i}{2}\int_{-\infty}^{\infty} dk_\rho \frac{k_\rho}{k_z} H_0^{(1)}(k_\rho\rho)e^{ik_z|z|}. \tag{35.1.15}$$

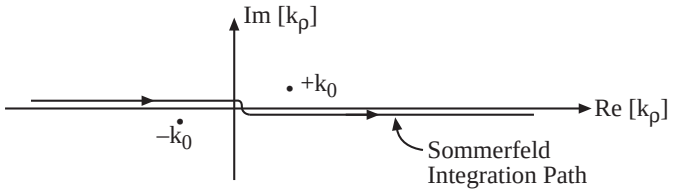


Figure 35.4: This figure illustrates the Sommerfeld integration path. Along this path, the integrand is well defined, as the path does not collide with any of the singularities of the integrand. Due to the existence of branch cuts and branch points, the integration path has to be defined on the correct Riemann sheet (see discussion in the text).

Since $H_0^{(1)}(x)$ has a logarithmic branch-point singularity at $x = 0$,[10] and $k_z = (k_0^2 - k_\rho^2)^{1/2}$ has algebraic branch-point singularities at $k_\rho = \pm k_0$, the integrand is a multi-value function. For proper book-keeping of a multi-value functions, branch cuts and Riemann sheets need to be defined [92, 108]. Thus, the integral in Equation (35.1.15) is undefined unless we stipulate also the path of integration and the Riemann sheet. Thus, a path of integration adopted by Sommerfeld, which is even good for a lossless medium, is shown in Figure 35.4. Because of the manner in which we have selected the reflection formula for Hankel functions, i.e., $H_0^{(1)}(e^{i\pi}x) = -H_0^{(2)}(x)$, the path of integration should be above the logarithmic branch-point singularity at the origin. With this definition of the Sommerfeld integration, the integral is well defined even when there is no loss, i.e., when the branch points $\pm k_0$ are on the real axis.

35.2 A Source on Top of a Layered Medium

Previously, we have studied the propagation of plane electromagnetic waves from a single dielectric interface in Section 18.1 as well as through a layered medium in Section 20.1. It can be shown that plane waves reflecting from a layered medium can be decomposed into TE-type plane waves, where $E_z = 0$, $H_z \neq 0$, and TM-type plane waves, where $H_z = 0$, $E_z \neq 0$.[11] One also sees how the field due to a point source can be expanded into plane waves in Section 35.1.

In view of the above observations, when a point source is on top of a layered medium, it is then best to decompose the field of the point source in terms of plane waves of TE type and TM type.

[9] The reflection formula allows the analytic continuation of the integrand from the positive k_ρ to the negative k_ρ axis.

[10] $H_0^{(1)}(x) \sim \frac{2i}{\pi}\ln(x)$, see Chew [37][p. 14], or Abromawitz or Stegun [113].

[11] *Chew, Waves and Fields in Inhomogeneous Media* [37]; *Kong, Electromagnetic Wave Theory* [34].

Then, the nonzero component of E_z characterizes TM-to-z waves, while the nonzero component of H_z characterizes TE-to-z waves. Hence, given a field, its TM and TE components can be extracted readily. Furthermore, if these TM and TE components are expanded in terms of plane waves, their propagations in a layered medium can be studied easily.

The problem of a vertical electric dipole on top of a half space was first solved by Sommerfeld (1909) [155] using Hertzian potentials, which are related to the z components of the electromagnetic field. The work is later generalized to layered media, as discussed in the literature. Later, Kong (1972) [262] suggested the use of the z components of the electromagnetic field instead of the Hertzian potentials.

35.2.1 Electric Dipole Fields–Spectral Expansion

The representation of a spherical wave in terms of plane waves can be done using Weyl identity or Sommerfeld identiy. Here, we will use Sommerfeld identity in anticipation of simpler numerical integration, since only single integrals are involved. The $\mathbf{E}$ field in a homogeneous medium due to a point current source or a Hertzian dipole directed in the $\hat{\alpha}$ direction, $\mathbf{J} = \hat{\alpha} I\ell\, \delta(\mathbf{r})$, is derivable via the vector potential method or the dyadic Green's function approach. Then, using the dyadic Green's function approach, or the vector/scalar potential approach, the field due to a Hertzian dipole is given by

$$\mathbf{E}(\mathbf{r}) = i\omega\mu \left(\bar{\mathbf{I}} + \frac{\nabla\nabla}{k^2} \right) \cdot \hat{\alpha} I\ell\, \frac{e^{ikr}}{4\pi r}, \tag{35.2.1}$$

where $I\ell$ is the current moment and $k = \omega\sqrt{\mu\epsilon}$, the wave number of the homogeneous medium. Furthermore, from $\nabla \times \mathbf{E} = i\omega\mu\mathbf{H}$, the magnetic field due to a Hertzian dipole is shown to be given by (after using $\nabla \times \nabla = 0$)

$$\mathbf{H}(\mathbf{r}) = \nabla \times \hat{\alpha} I\ell\, \frac{e^{ikr}}{4\pi r}. \tag{35.2.2}$$

With the above fields, their TM-to-z and TE-to-z components can be extracted easily in anticipation of their plane wave expansions for propagation through layered media.

(a) Vertical Electric Dipole (VED)—Spectral Expansion

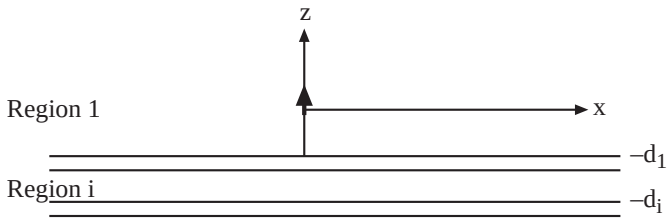


Figure 35.5: A vertical electric dipole over a layered medium.

We start with the vertical electric dipole first because it is simpler. A vertical electric dipole shown in Figure 35.5 has $\hat{\alpha} = \hat{z}$; hence, in anticipation of their plane wave expansions, the TM-to-z component of the field is characterized by $E_z \neq 0$ or that

$$E_z = \frac{i\omega\mu I\ell}{4\pi k^2}\left(k^2 + \frac{\partial^2}{\partial z^2}\right)\frac{e^{ikr}}{r}, \tag{35.2.3}$$

and the TE component of the field is characterized by

$$H_z = 0, \tag{35.2.4}$$

implying the absence of the TE-to-z field.

Next, using the Sommerfeld identity (35.1.15) in the above, and after exchanging the order of integration and differentiation, we have[12]

$$E_z = \frac{-I\ell}{4\pi\omega\epsilon}\int_0^\infty dk_\rho \frac{k_\rho^3}{k_z} J_0(k_\rho\rho)e^{ik_z|z|},\ |z| \neq 0 \tag{35.2.5}$$

after noting that $k_\rho^2 + k_z^2 = k^2$. Notice that now Equation (35.2.5) expands the z component of the electric field in terms of cylindrical waves in the ρ direction and a plane wave in the z direction. (Cylindrical waves actually are linear superpositions of plane waves, because we can work backward from (35.1.15) to (35.1.11) to see this.) As such, the integrand in (35.2.5) in fact consists of a linear superposition of TM-type plane waves. The above is also the *primary field* generated by the source.[13]

Consequently, for a VED on top of a stratified medium as shown, expanding the source field in terms of plane waves, the downgoing plane waves from the point source will be reflected like TM waves with the generalized reflection coefficient $\tilde{R}_{12}^{TM}$. Henceforth, over a stratified medium, the field in region 1 can be written as

$$E_{1z} = \frac{-I\ell}{4\pi\omega\epsilon_1}\int_0^\infty dk_\rho \frac{k_\rho^3}{k_{1z}} J_0(k_\rho\rho)\left[e^{ik_{1z}|z|} + \tilde{R}_{12}^{TM}e^{ik_{1z}z+2ik_{1z}d_1}\right], \tag{35.2.6}$$

where $k_{1z} = (k_1^2 - k_\rho^2)^{\frac{1}{2}}$, and $k_1^2 = \omega^2\mu_1\epsilon_1$, the wave number in region 1.

The phase-matching condition dictates that the transverse variation of the field in all the regions must be the same. Consequently, in the i-th region, the solution becomes[14]

$$\epsilon_i E_{iz} = \frac{-I\ell}{4\pi\omega}\int_0^\infty dk_\rho \frac{k_\rho^3}{k_{1z}} J_0(k_\rho\rho)A_i\left[e^{-ik_{iz}z} + \tilde{R}_{i,i+1}^{TM}e^{ik_{iz}z+2ik_{iz}d_i}\right]. \tag{35.2.7}$$

[12]By using (35.1.15) in (35.2.3), the $\partial^2/\partial z^2$ operating on $e^{ik_z|z|}$ produces a Dirac delta function singularity. But for simplicity, in (35.2.5), we ignore the delta function since $|z| \neq 0$. Hence, $\partial^2/\partial z^2$ produces a $-k_z^2$, and $k^2 - k_z^2 = k_\rho^2$. Detail discussion on this can be found in the chapter on dyadic Green's function in *Chew, Waves and Fields in Inhomogeneous Media* [37].

[13]One can perform a sanity check on the odd and even symmetry of the fields' z-component by sketching the fields of a static vertical electric dipole.

[14]It will take quite a bit of work to get this expression, but you just need to know that it can be done, and know where to look for the resources for it.

Notice that Equation (35.2.7) is now expressed in terms of $\epsilon_i E_{iz}$ because $\epsilon_i E_{iz}$ reflects and transmits like H_{iy}, the transverse component of the magnetic field or TM waves.[15] Therefore, $\tilde{R}^{TM}_{i,i+1}$ and A_i could be obtained using the methods discussed in *Chew, Waves and Fields in Inhomogeneous Media* [108].

This completes the derivation of the integral representation of the electric field everywhere in the stratified medium. These integrals are known as ***Sommerfeld integrals***. The case when the source is embedded in a layered medium can be derived similarly.

(b) Horizontal Electric Dipole (HED)—Spectral Expansions

The HED is more complicated. Unlike the VED that excites only the TM-to-z waves, an HED will excite both TE-to-z and TM-to-z waves. For a horizontal electric dipole pointing in the x direction, $\hat{\alpha} = \hat{x}$; hence, (35.2.1) and (35.2.2) give the TM-to-z and the TE-to-z components, in anticipation of their plane wave expansions, as

$$E_z = \frac{iI\ell}{4\pi\omega\epsilon}\frac{\partial^2}{\partial z\partial x}\frac{e^{ikr}}{r}, \tag{35.2.8}$$

$$H_z = -\frac{I\ell}{4\pi}\frac{\partial}{\partial y}\frac{e^{ikr}}{r}. \tag{35.2.9}$$

Then, with the Sommerfeld identity (35.1.15), we can expand the above as (after exchanging the order of differentiation and integration)

$$E_z = \pm\frac{iI\ell}{4\pi\omega\epsilon}\cos\phi \int_0^\infty dk_\rho\, k_\rho^2 J_1(k_\rho\rho)e^{ik_z|z|} \tag{35.2.10}$$

$$H_z = i\frac{I\ell}{4\pi}\sin\phi \int_0^\infty dk_\rho\, \frac{k_\rho^2}{k_z} J_1(k_\rho\rho)e^{ik_z|z|}. \tag{35.2.11}$$

The $\pm$ sign above comes from the ∂_z derivative involving $|z|$. It also indicates that E_z is odd symmetric about $z = 0$, which one can easily verify by sketching the field of a horizontal electric dipole. Now, Equation (35.2.10) represents the wave expansion of the TM-to-z field, while (35.2.11) represents the wave expansion of the TE-to-z field in terms of Sommerfeld integrals which are plane-wave expansions in disguise. Observe that because E_z is odd about $z = 0$ in (35.2.10), the downgoing wave has an opposite sign from the upgoing wave. At this point, the above are just the primary field generated by the source.

On top of a stratified medium, the downgoing wave is reflected accordingly, depending on its

[15] See *Chew, Waves and Fields in Inhomogeneous Media* [37], p. 46, (2.1.6) and (2.1.7). Or we can gather from (18.1.6) to (18.1.7) that the $\mu_i H_{iz}$ transmits like E_{iy} at a dielectric interface, and by duality, $\epsilon_i E_{iz}$ transmits like H_{iy}.

wave type. Consequently, we have

$$E_{1z} = \frac{iI\ell}{4\pi\omega\epsilon_1}\cos\phi \int_0^\infty dk_\rho\, k_\rho^2 J_1(k_\rho\rho)\left[\pm e^{ik_{1z}|z|} - \tilde{R}_{12}^{TM} e^{ik_{1z}(z+2d_1)}\right], \tag{35.2.12}$$

$$H_{1z} = \frac{iI\ell}{4\pi}\sin\phi \int_0^\infty dk_\rho\, \frac{k_\rho^2}{k_{1z}} J_1(k_\rho\rho)\left[e^{ik_{1z}|z|} + \tilde{R}_{12}^{TE} e^{ik_{1z}(z+2d_1)}\right]. \tag{35.2.13}$$

Notice that the negative sign in front of $\tilde{R}_{12}^{TM}$ in (35.2.12) follows because the downgoing wave in the primary field has a negative sign as shown in (35.2.10).

In the Sommerfeld integrals, $k_{iz} = \sqrt{k_i^2 - k_\rho^2}$ are double-value functions. So the integrands are potentially multi-value functions and care has to be taken when evaluating these integrals. Proper book-keeping is done by defining branch cuts and branch points. However, it can be shown that, sometimes, these double-value functions do not make the integrand double value, and this is discussed in [37, p. 112].

35.3 Stationary Phase Method and Fermat's Principle

Sommerfeld integrals are rather complex, and by themselves, they do not offer much insight into the physics of the field. To elucidate the physics, we can apply the stationary phase method to approximate these integrals when the frequency is high, or when kr is large, or when the observation point is many wavelengths away from the source point. It turns out that this method is intimately related to Fermat's principle described in Section 33.2.

Stationary Phase Method (SPM)–A Canonical Case

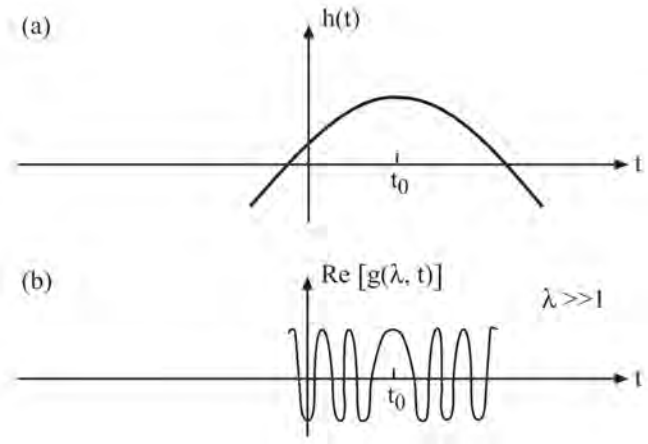


Figure 35.6: We first look at $h(t)$ which is in the exponent of $g(\lambda, t)$ (see (35.3.2)). Here, $h(t)$ is a slowly varying function of t as shown in (a). But it is in the exponent multiplied by λ. When λ is large, the function $g(\lambda, t)$ is rapidly varying as shown in (b).

SPM is useful in deriving the leading-order approximation to an integral whose integrand is rapidly oscillating. Such rapid oscillations render the numerical integration of the integral difficult. In this case, an asymptotic approximation is the best way to obtain numerical values for the integral.

As an example, consider a basic, canonical integral of the form

$$I = \int_{-\infty}^{\infty} dt\, f(t)\, g(\lambda, t) \tag{35.3.1}$$

where $g(\lambda, t)$ can be integrated in closed form, but not $f(t)g(\lambda, t)$. Furthermore, if

$$g(\lambda, t) \sim e^{i\lambda h(t)}, \quad \lambda \to \infty \tag{35.3.2}$$

("$\sim$" means "asymptotic to") then when λ is large, $g(\lambda, t)$ is a rapidly oscillating function of h, and hence, of t, where $h(t)$ may be a function of t as shown in Figure 35.6(a). As a result, the real part of the function $g(\lambda, t)$ will look like that shown in Figure 35.6b. Also, the imaginary part of $g(\lambda, t)$ has the same feature. If $h(t)$ has a stationary point at $t = t_0$, i.e.,

$$h'(t_0) = 0 \tag{35.3.3}$$

it varies least slowly as a function of t at $t = t_0$; hence, $g(\lambda, t)$ is least rapidly oscillating at $t = t_0$. Furthermore, we assume that when $\lambda \to \infty$, $f(t)$ is a more slowly varying function compared to $g(\lambda, t)$.

If the integrand is rapidly oscillating, the contribution to the integration is small because of the cancellation of the positive and negative parts of the integrand. However, this cancellation is least

at t_0 where the integrand is least rapidly oscillating. Therefore, most of the contribution to the integration will come from the neighborhood of t_0. Consequently, the integral can be approximated as (see Chew [263])

$$I \sim f(t_0) \int_{-\infty}^{\infty} dt\, g(\lambda, t), \;\; \lambda \to \infty \tag{35.3.4}$$

because $f(t)$ is approximately a constant $f(t_0)$ in the neighborhood of t_0. Moreover, since $g(\lambda, t)$ has a closed-form solution, a simple algebraic approximation is obtained for I. This approximation is asymptotic in the sense that it becomes better when λ becomes larger.

Application of SPM to Sommerfeld Integrals

Next, we prepare the Sommerfeld integrals for SPM approximation. In order to avoid having to work with special functions like Bessel functions, we convert the Sommerfeld integrals back to spectral integrals in the cartesian coordinates. (We could have obtained the aforementioned integrals in cartesian coordinates were we to start with the Weyl identity instead of the Sommerfeld identity.) To do the back conversion, we make use of the identity,

$$\frac{e^{ik_0 r}}{r} = \frac{i}{2\pi} \iint_{-\infty}^{\infty} dk_x dk_y \, \frac{e^{ik_x x + ik_y y + ik_z |z|}}{k_z} = i \int_0^{\infty} dk_\rho \, \frac{k_\rho}{k_z} J_0(k_\rho \rho) e^{ik_z |z|}. \tag{35.3.5}$$

We can just focus our attention on the reflected wave term in (35.2.6) and rewrite it in cartesian coordinates to get

$$\begin{aligned} E_{1z}^R &= \frac{-I\ell}{8\pi^2 \omega \epsilon_1} \iint_{-\infty}^{\infty} dk_x dk_y \frac{k_x^2 + k_y^2}{k_{1z}} R_{12}^{TM} e^{ik_x x + ik_y y + ik_{1z}(z+2d_1)} \\ &= \iint_{-\infty}^{\infty} dk_x dk_y \frac{1}{k_{1z}} F(k_x, k_y) e^{ik_x x + ik_y y + ik_{1z}(z+2d_1)} \end{aligned} \tag{35.3.6}$$

where we have put all the complicated terms of the integrand in the function $F(k_x, k_y)$ defined as

$$F(k_x, k_y) = \frac{-I\ell}{8\pi^2 \omega \epsilon_1} (k_x^2 + k_y^2) R_{12}^{TM}$$

In the above, $k_x^2 + k_y^2 + k_{1z}^2 = k_1^2$ is the dispersion relation satisfied by the plane wave in region 1. Also, R_{12}^{TM} is dependent on $k_{iz} = \sqrt{k_i^2 - k_x^2 - k_y^2}$ in cartesian coordinates, where $i = 1, 2$. For simplicity, we will assume that $d_1 = 0$ to begin. Now the problem reduces to finding the approximation of the following integral:

$$E_{1z}^R = \iint_{-\infty}^{\infty} dk_x dk_y \frac{1}{k_{1z}} F(k_x, k_y) e^{irh(k_x, k_y)} \tag{35.3.7}$$

where

$$rh(k_x, k_y) = r\left(k_x \frac{x}{r} + k_y \frac{y}{r} + k_{1z} \frac{z}{r}\right) \tag{35.3.8}$$

The ratios x/r, y/r, and z/r are deliberately used because the are at most of $O(1)$ when $r \to \infty$. The large parameter here is $\lambda = r$. The above integral is too complicated to see its physics clearly. To elucidate its physics, we want to approximate the above integral when $rh(k_x, k_y)$ is large. This happens when x, y, and z are large compared to wavelength.

In the above, $e^{irh(k_x,k_y)}$ is a rapidly varying function of k_x and k_y when x, y, and z are large, or r is large compared to wavelength.[16] In other words, a small change in k_x or k_y will cause a large change in the phase of the integrand, or the integrand will be a rapidly varying function of k_x and k_y. Due to the cancellation of the integral when one integrates a rapidly varying function, most of the contributions to the integral will come from around the stationary point of $h(k_x, k_y)$ or where the function is least slowly varying (see Figure 35.6).[17] Otherwise, the integrand is rapidly varying away from this point, and the integration contributions will destructively cancel with each other, while around the stationary point, they will add constructively.

The stationary point of $h(k_x, k_y)$ in the k_x and k_y plane is found by setting the derivatives of $h(k_x, k_y)$ with respect to k_x and k_y to zero. By so doing

$$\frac{\partial h}{\partial k_x} = \frac{x}{r} - \frac{k_x}{k_{1z}} \frac{z}{r} = 0, \qquad \frac{\partial h}{\partial k_y} = \frac{y}{r} - \frac{k_y}{k_{1z}} \frac{z}{r} = 0 \tag{35.3.9}$$

The above represents two equations from which the two unknowns, k_{xs} and k_{ys}, at the stationary phase point can be solved for. By expressing the above in spherical coordinates, $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$, the values of (k_{xs}, k_{ys}), that satisfy the above equations are

$$k_{xs} = k_1 \sin\theta\cos\phi, \quad k_{ys} = k_1 \sin\theta\sin\phi \tag{35.3.10}$$

with the corresponding $k_{1zs} = k_1\cos\theta$. Here, the wave vector $\mathbf{k}_s = \hat{x}k_{xs} + \hat{y}k_{ys} + \hat{z}k_{zs}$ is parallel to the vector that connects that source and the observation point, viz., $\mathbf{r} = \hat{x}x + \hat{y}y + \hat{z}z$. In this case, $\mathbf{k}_s \cdot \mathbf{r} = kr$.

[16] The yardstick in wave physics is always wavelength. Large distance is also synonymous to increasing the frequency or reducing the wavelength.

[17] We have replaced a 1D function $h(t)$ with a 2D function $h(k_x, k_y)$, but the idea is the same.

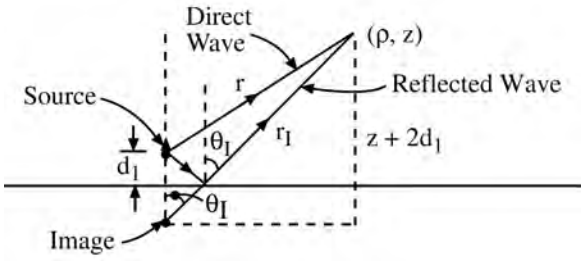


Figure 35.7: At high frequencies, the source point and the observation point are connected by a ray. The ray represents a bundle of plane waves that interfere constructively. This is even true for a bundle of plane waves that reflect off an interface. So ray theory or ray optics prevails here, and the ray bounces off the interface according to the reflection coefficient of a plane wave impinging at the interface with θ_I.

When one integrates on the k_x and k_y plane, the dominant contribution to the integral will come from the point in the vicinity of (k_{xs}, k_{ys}). Assuming that $F(k_x, k_y)$ is slowly varying, we can approximate $F(k_x, k_y)$ to be a constant equal to its value at the stationary phase point (k_{xs}, k_{ys}), and say that[18]

$$E^R_{1z} \simeq F(k_{xs}, k_{ys}) \iint_{-\infty}^{\infty} \frac{1}{k_{1z}} e^{ik_x x + ik_y y + ik_{1z} z} dk_x dk_y = 2\pi F(k_{xs}, k_{ys}) \frac{e^{ik_1 r}}{ir} \tag{35.3.11}$$

In the above, the integral can be performed in closed form using the Weyl identity. The physical meaning of the above is that if we begin with a complicated integral such as that in (35.3.7), which represents an integral summation of plane waves in all directions, it can be approximated by a spherical wave. The above analysis says that only plane waves in the vicinity of the stationary point contribute to the integral, and these plane waves are paraxial. Thus a spherical wave can be thought of as a summation of a bundle of paraxial plane waves.

When $d_1 \neq 0$, we will have to replace (35.3.8) by

$$r_I h(k_x, k_y) = r_I \left(k_x \frac{x}{r_I} + k_y \frac{y}{r_I} + k_{1z} \frac{z + 2d_1}{r_I} \right) \tag{35.3.12}$$

where $r_I = \sqrt{x^2 + y^2 + (z + 2d_1)^2}$ is the distance from the image point to the observation point (see Figure 35.7). The problem is 'homomorphic' to the previous case. By expressing the above in spherical coordinates whose origin is the image point, $x = r_I \sin\theta_I \cos\phi$, $y = r_I \sin\theta_I \sin\phi$, $z + 2d_1 = r_I \cos\theta_I$, the values of (k_{xs}, k_{ys}), that satisfy the above equations are

$$k_{xs} = k_1 \sin\theta_I \cos\phi, \quad k_{ys} = k_1 \sin\theta_I \sin\phi \tag{35.3.13}$$

[18]The rapidly varying function with a stationary phase point is behaving like the sifting property of a delta function.

The subsequent stationary phase approximation yields

$$E_{1z}^R \simeq F(k_{xs}, k_{ys}) \iint_{-\infty}^{\infty} \frac{1}{k_{1z}} e^{ik_x x + ik_y y + ik_{1z}(z+2d_1)} dk_x dk_y = 2\pi F(k_{xs}, k_{ys}) \frac{e^{ik_1 r_I}}{ir_I} \tag{35.3.14}$$

The above expressions have a few important physical interpretations.

(i) Even though a source is emanating plane waves in all directions in accordance to (35.1.11), at the observation point r or r_I far away from the image source point, only one or few plane waves in the vicinity of the stationary phase point are important. They interfere with each other constructively to form a spherical wave that represents the ray connecting the source point to the observation point. Plane waves in other directions interfere with each other destructively, and are not important. That is the reason that the source point and the observation point are connected only by one ray, or one bundle of plane waves in the vicinity of the stationary phase point. These bundle of plane waves are also almost paraxial with respect to each other. This yields the insight that a ray is a bundle of plane waves who are mostly paraxial with respect to each other! They add coherently to form a spherical wave front.

(ii) The function $F(k_x, k_y)$ could be a very complicated function like the reflection coefficient R^{TM}, but only its value at the stationary phase point matters the most. If we were to make $d_1 \neq 0$, the math remains similar except that now, we replace r with $r_I = \sqrt{x^2 + y^2 + (z + 2d_1)^2}$, the distance from the observation point to the image source point. The reason being that due to the reflecting half-space, the source point now has an image source point as shown in Figure 35.7. Also, we have to be mindful to replace z with $z + 2d_1$ in the above stationary phase analysis. The stationary phase method now extract a ray (or a bundle of plane waves) that emanates from the source point, bounces off the half-space, and the reflected ray reaches the observer modulated by the reflection coefficient R^{TM}. But the value of the reflection coefficient that matters the most is at the angle at which the incident ray impinges on the half-space. Now, the reflected field seems to have emanated from an image source point.

(iii) At the stationary phase point, the ray is formed by the $\mathbf{k}_I$-vector where

$$\mathbf{k}_I = \hat{x} k_1 \sin\theta_I \cos\phi + \hat{y} k_1 \sin\theta_I \sin\phi + \hat{z} k_1 \cos\theta_I$$

This is a $\mathbf{k}_I$ vector that points from the image source point to the observation point. This ray points in the same direction as the position vector of the observation point with respect to the image source point, viz., $\mathbf{r}_I = \hat{x} r_I \sin\theta_I \cos\phi + \hat{y} r_I \sin\theta_I \sin\phi + \hat{z} r_I \cos\theta_I$. In other words, the $\mathbf{k}_I$-vector and the $\mathbf{r}_I$-vector point in the same direction. This is reminiscent of Fermat principle, because when this happens, the ray propagates with the minimum phase shift and time delay between the image source point and the observation point, which is also the statement of Fermat's principle. When $z \to z + 2d_1$, the ray for the image source is as shown in Figure 35.7 where the ray is minimum phase from the image source to the observation point. Hence, the stationary phase method is intimately related to Fermat's principle.

Exercises for Lecture 35

Problem 35-1: This exercise refers to Chapter 35 of the lecture notes.

(i) By taking the residue of a contour integral around a pole, show how you can go from eq. (35.1.7) to eq. (35.1.9) and eq. (35.1.10).

(ii) Show that eq. (35.2.3) can be derived from (35.2.1), and then show that (35.2.5) can be derived assuming that $z \neq 0$. Similarly, from (35.2.1), derive (35.2.8) and (35.2.9), and then show that (35.2.10) and (35.2.11) can be derived.

(iii) First, for the section on stationary phase, derive (35.3.6) which is the reflected field term in Cartesian coordinates. Assume that $d_1 \neq 0$, derive the equivalence of (35.3.14) with $r \to r_I = \sqrt{x^2 + y^2 + (z + 2d_1)^2}$, using stationary phase argument, and elucidate the physics expressed in the math in accordance to item (ii) below (35.3.14). Explain why this is related to Fermat's principle.

Chapter 36

Computational Electromagnetics, Numerical Methods

Due to the rapid advent of digital computers and the blinding speed at which computations can be done, numerical methods to seek solutions of Maxwell's equations have become vastly popular. Massively parallel digital computers now can compute at breakneck speed of tera\peta\exa-flops throughputs [264], where FLOPS stands for "floating operations per second". They have also spawn terms that we have not previously heard of (see also Figure 36.1).

Computer performance

Name	Unit	Value
kiloFLOPS	kFLOPS	10^3
megaFLOPS	MFLOPS	10^6
gigaFLOPS	GFLOPS	10^9
teraFLOPS	TFLOPS	10^{12}
petaFLOPS	PFLOPS	10^{15}
exaFLOPS	EFLOPS	10^{18}
zettaFLOPS	ZFLOPS	10^{21}
yottaFLOPS	YFLOPS	10^{24}

Figure 36.1: Nomenclature for measuring the speed of modern day computers. The fastest computer in the world changes from year to year. It is updated on this website [265]

We repeat a quote from Freeman Dyson—"Technology is a gift of God. After the gift of life it is perhaps the greatest of God's gifts. It is the mother of civilizations, of arts and of sciences." The spur for computer advancement is due to the second world war. During then, men went to war while women stayed back to work as computers, doing laborious numerical computations manually

(see Figure 36.2 [266]): The need for a faster computer is obvious. (Trajectories of rockets cannot be solved for in closed-form: computers are needed.) Unfortunately, in the last half century or so, we have been using a large part of the gift of technology in warfare to destroy God's greatest gift, life!

Figure 36.2: A woman working as a "computer" shortly after the second world war (courtesy of Wikipedia [266]).

36.1 Computational Electromagnetics, Numerical Methods

Due to the high fidelity of Maxwell's equations in describing electromagnetic physics in nature, and they have been validated to high accuracy (see Section 1.1), often time, a numerical solution obtained by solving Maxwell's equations is more reliable than laboratory experiments. This field of finding numerical solutions to Maxwell's equations is also known as ***computational electromagnetics*** (CEM).

The field is a descendent of ***m***athematical modeling. It begins with the quest for closed form solutions or analytic solutions that can return a number with some simple calculations. These closed form solutions have greatly benefitted engineering designs. Some examples of these are the Mie scattering solution, the Sommerfeld half-space problems, and many more. As shall be learnt from this chapter, many wave scattering problems can be couched in terms of solving a matrix equation. For many highly complex problems, a high fidelity description of the problem can be gotten using many degrees of freedom, or the increasing the number of unknowns in the pertinent matrix equations. Numerical methods exploit the blinding speed of modern digital computers to perform calculations, and hence to solve large matrix system of equations. This manner of solving engineering problems is also termed numerical simulation. It turns out that the design of a modern-day computer chip is done with numerical simulation more than 90 percent of the time.

Computational electromagnetics consists mainly of two classes of numerical solvers: one that solves differential equations directly: the differential-equation solvers; and one that solves integral equations: the integral equation solvers. Both these classes of equations are derivable from Maxwell's equations.[1]

[1] Computations are heavily used in other fields such as computational mechanics, computational fluid dynamics,

36.2 Examples of Differential Equations

An example of differential equation with no closed form solution, but driven by a source is the scalar wave equation:

$$(\nabla^2 + k^2(\mathbf{r}))\,\phi(\mathbf{r}) = Q(\mathbf{r}), \tag{36.2.1}$$

An example of vector differential equation for vector electromagnetic field is

$$\nabla \times \overline{\boldsymbol{\mu}}^{-1} \cdot \nabla \times \mathbf{E}(\mathbf{r}) - \omega^2 \overline{\boldsymbol{\epsilon}}(\mathbf{r}) \cdot \mathbf{E}(\mathbf{r}) = i\omega \mathbf{J}(\mathbf{r}) \tag{36.2.2}$$

These equations are linear equations,[2] but for inhomogeneous media where $k^2(\mathbf{r})$ and $\epsilon(\mathbf{r})$ are functions of position vector $\mathbf{r}$, generally, they do not have closed form solutions: then numerical solution is the norm. These problems above expressed by (36.2.1) and (36.2.2) share one commonality, i.e., they can be abstractly written as

$$\mathscr{L} f = g \tag{36.2.3}$$

where $\mathscr{L}$ is the differential operator which is linear,[3] and f is the unknown, and g is the driving source. Differential equations, or partial differential equations, as mentioned before in (36.2.1) and (36.2.2), have to be solved with boundary conditions. Otherwise, there is no unique solution to these equations.

In the case of the scalar wave equation (36.2.1), $\mathscr{L} = (\nabla^2 + k^2)$ is a differential operator. In the case of the electromagnetic vector wave equation (36.2.2), $\mathscr{L} = (\nabla \times \overline{\boldsymbol{\mu}}^{-1} \cdot \nabla \times) - \omega^2 \overline{\boldsymbol{\epsilon}} \cdot$. Furthermore, f will be $\phi(\mathbf{r})$ for the scalar wave equation (36.2.1), while it will be $\mathbf{E}(\mathbf{r})$ in the case of vector wave equation for an electromagnetic system (36.2.2). The g on the right-hand side can represent Q in (36.2.1) or $i\omega\mathbf{J}(\mathbf{r})$ in (36.2.2).

---

as well as computational physics.

[2] Nonlinear equations are often approximated as a series of linear problems.

[3] The simple test of linearity is that $\mathscr{L}(a_1 f_1 + a_2 f_2) = a_1 \mathscr{L} f_1 + a_2 \mathscr{L} f_2$. This test is homomorphic to the linearity test in linear algebra.

36.3 Examples of Integral Equations

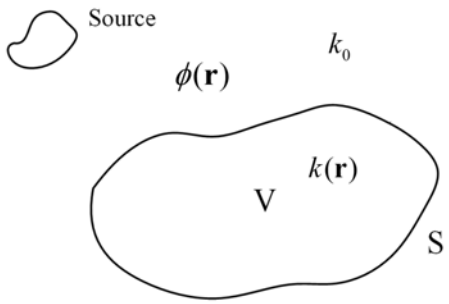


Figure 36.3: Geometry for the derivation of the volume-integral equation for scalar waves. The wavenumber $k(\mathbf{r})$ is assumed to be inhomogeneous, and hence, a function of position $\mathbf{r}$ inside the scatterer, but a constant k_0 outside the scatterer. The dyadic Green's function or vector case is homomorphic to the scalar wave case.

36.3.1 Volume Integral Equation

This course is replete with PDE's, but we have not come across too many integral equations as yet. In integral equations, the unknown is embedded in the integral. The simplest integral equation to derive is the volume integral equation. Hence, we shall first derive the volume integral equation for the scalar wave case.[4] In this case, the pertinent scalar wave equation is

$$[\nabla^2 + k^2(\mathbf{r})]\phi(\mathbf{r}) = Q(\mathbf{r}), \tag{36.3.1}$$

where $k^2(\mathbf{r})$ represents an inhomogeneous medium over a finite domain V, and $k^2 = k_0^2$, which is constant outside V (see Figure 36.3). Next, we define a Green's function satisfying

$$[\nabla^2 + k_0^2]g(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}'), \qquad \forall \mathbf{r}, \mathbf{r}'. \tag{36.3.2}$$

Then, (36.3.1) can be rewritten as

$$[\nabla^2 + k_0^2]\phi(\mathbf{r}) = Q(\mathbf{r}) - [k^2(\mathbf{r}) - k_0^2]\phi(\mathbf{r}). \tag{36.3.3}$$

Note that the right-hand side of (36.3.3) can be considered an equivalent source. Since the Green's function corresponding to the differential operator on the left-hand side of (36.3.3) is known, by the principle of linear superposition, we can write the formal solution to (36.3.3) as

$$\phi(\mathbf{r}) = -\int_{V_s} dV' g(\mathbf{r}, \mathbf{r}')Q(\mathbf{r}') + \int_V dV' g(\mathbf{r}, \mathbf{r}')[k^2(\mathbf{r}') - k_0^2]\phi(\mathbf{r}'). \tag{36.3.4}$$

[4]The vector wave case is homomorphic to the scalar wave case.

The first term on the right-hand side is just the field due to the source in the absence of the inhomogeneity or the scatterer, and hence, is the incident field. The second term is a volume integral over the space where $k^2(\mathbf{r}') - k_0^2 \neq 0$, or inside the inhomogeneous scatterer. Therefore, (36.3.4) becomes

$$\phi(\mathbf{r}) = \phi_{inc}(\mathbf{r}) + \int_V dV' g(\mathbf{r}, \mathbf{r}')[k^2(\mathbf{r}') - k_0^2]\phi(\mathbf{r}'). \tag{36.3.5}$$

It is to be noted that the above sources are radiating via the Green's function, and hence they satisfy the radiation condition, since the Green's function satisfies the radiation condition.

In the above equation, if the total field $\phi(\mathbf{r}')$ inside the volume V on the right-hand side is known, then $\phi(\mathbf{r})$ can be calculated everywhere. But $\phi(\mathbf{r})$ is unknown at this point. To solve for $\phi(\mathbf{r})$, an integral equation has to be formulated for $\phi(\mathbf{r})$. To this end, we imposed (36.3.5) for $\mathbf{r}$ in V. Then, $\phi(\mathbf{r})$ on the left-hand side and on the right-hand side are the same unknown defined over the same region V. Consequently, (36.3.5) becomes the desired integral equation after rearrangement as

$$\phi_{inc}(\mathbf{r}) = \phi(\mathbf{r}) - \int_V dV' g(\mathbf{r}, \mathbf{r}')[k^2(\mathbf{r}') - k_0^2]\phi(\mathbf{r}'), \quad \mathbf{r} \in V. \tag{36.3.6}$$

In the above, the unknown $\phi(\mathbf{r})$ is defined over a volume V, over which the integration is performed, and hence the name volume integral equation. Alternatively, the above can be rewritten, using shorthand notation, as

$$\phi_{inc}(\mathbf{r}) = \phi(\mathbf{r}) - \mathcal{G}(\mathbf{r}, \mathbf{r}')\mathcal{O}(\mathbf{r}')\phi(\mathbf{r}'), \quad \mathbf{r} \in V, \tag{36.3.7}$$

where $\mathcal{G}$ is the integral operator in (36.3.6),[5] and $\mathcal{O}(\mathbf{r}') = [k^2(\mathbf{r}') - k_0^2]$ is the scatterer object function. It is also a *Fredholm integral equation* of the second kind because the unknown is both inside and outside the integral operator. In the above, integration over repeated variable $\mathbf{r}'$ is implied. Nevertheless, it can be written more abstractly as

$$\mathscr{L} f = g \tag{36.3.8}$$

where $\mathscr{L}$ is a linear operator ("homomorphic" to a linear matrix operator), while f represents the unknown function $\phi(\mathbf{r})$ and g is the known function $\phi_{inc}(\mathbf{r})$ (f and g are homomorphic to matrix vectors). In the above

$$\mathscr{L} = \mathcal{I} - \mathcal{G}\mathcal{O}, \qquad f = \phi(\mathbf{r}), \qquad g(\mathbf{r}) = \phi(\mathbf{r}) \tag{36.3.9}$$

where $\mathcal{I}$ is the identity operator.

[5] Sometimes, this is called the kernel of the integral equation.

36.3.2 Surface Integral Equation

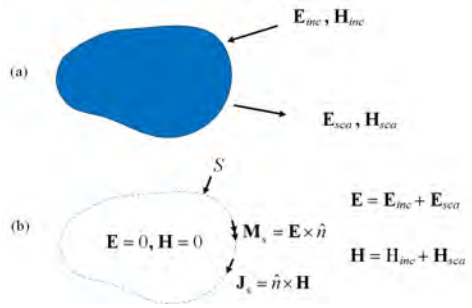


Figure 36.4: Geometry for the derivation of the surface-integral equation for vector electromagnetics waves. (a) The original electromagnetics scattering problem. (b) The equivalent electromagnetics problem by invoking equivalence principle.

The surface integral equation method is rather popular in many applications, because it can employ a homogeneous-medium Green's function which exists in simple closed-form,[6] and the unknowns reside on a surface rather than in a volume.[7]

The surface integral equation for vector electromagnetic field can be derived using the equivalence theorem also called the Love's equivalence theorem [105]. Given a scattering problem shown in Figure 36.4(a), it can be replaced by an equivalence problem as shown in Figure 36.4(b). One can verify this by performing a Gedanken experiment as we have done for the other equivalence problems discussed in Section 13.1.

In this figure, the total fields outside the scatterer are $\mathbf{E} = \mathbf{E}_{inc} + \mathbf{E}_{sca}$ and $\mathbf{H} = \mathbf{H}_{inc} + \mathbf{H}_{sca}$. The impressed equivalence currents are given by $\mathbf{M}_s = \mathbf{E} \times \hat{n}$, and $\mathbf{J}_s = \hat{n} \times \mathbf{H}$. These impressed currents, together generate the scattered fields outside the scatterer, while they generate zero field inside the scatterer!

As such, the scattered fields outside the scatterer can be found from the radiation of the impressed currents $\mathbf{M}_s$ and $\mathbf{J}_s$. Notice that these currents are radiating via the free-space Green's function because the scatterer has been removed in this equivalence problem. Now that if the scatterer is a PEC, then the tangential component of the total electric field is zero on the PEC surface. Therefore, $\mathbf{M}_s = 0$ and only $\mathbf{J}_s$ is radiating via the free-space Green's function.

Note that this equivalence problem is very different from that of an impressed currents on the PEC scatterer as discussed in Section 13.2. There, only the magnetic surface current is radiating in the presence of the PEC, and the Green's function is that of a current source radiating in the presence of the PEC scatterer, and it is not the free-space Green's function.

Now we can write the fields outside the scatterer using (13.4.20)

$$\mathbf{E}_{sca}(\mathbf{r}) = \frac{1}{i\omega\epsilon} \nabla \times \nabla \times \oint_S dS'\, g(\mathbf{r} - \mathbf{r}')\, \hat{n}' \times \mathbf{H}(\mathbf{r}') = \frac{1}{i\omega\epsilon} \nabla \times \nabla \times \oint_S dS'\, g(\mathbf{r} - \mathbf{r}')\, \mathbf{J}_s(\mathbf{r}') \qquad (36.3.10)$$

[6] Numerical Green's functions have been proposed to enable solutions of inhomogeneous media [267].

[7] These are sometimes called boundary integral equations method [268, 269].

In the above, we have swapped $\mathbf{r}'$ and $\mathbf{r}$ compared to (13.4.20). Also, we have kept only the electric current $\mathbf{J}_s(\mathbf{r})$ due to $\hat{n}\times\mathbf{H}(\mathbf{r})$. If we impose the boundary condition that the tangential component of the total electric field is zero, then we arrive at $\hat{n}\times\mathbf{E}_{sca} = -\hat{n}\times\mathbf{E}_{inc}$ and the integral equation

$$-\hat{n}\times\mathbf{E}_{inc}(\mathbf{r}) = \hat{n}\times\frac{1}{i\omega\epsilon}\nabla\times\nabla\times\oint_S dS'\, g(\mathbf{r}-\mathbf{r}')\,\mathbf{J}_s(\mathbf{r}'). \qquad \mathbf{r}\in S \tag{36.3.11}$$

In the above, $\hat{n}\times\mathbf{E}_{inc}(\mathbf{r})$ is known on the left hand side on the scatterer's surface, while the right-hand side has embedded in it the unknown surface current $\mathbf{J}_s(\mathbf{r}) = \hat{n}\times\mathbf{H}(\mathbf{r})$ on the surface the scatter. Therefore, the above is an integral equation for the unknown surface current $\mathbf{J}_s(\mathbf{r})$. It can be written as a form of $\mathscr{L}f = g$ just like other linear operator equations.

36.4 Function as a Vector

Several linear operator equations have been derived in the previous sections. They are all of the form

$$\mathscr{L}f = g \tag{36.4.1}$$

In the above, f is a functional vector which is the analogue of the vector $\mathbf{f}$ in matrix theory or linear algebra.[8] In linear algebra, the vector $\mathbf{f}$ is of length N in an N dimensional space. It can be indexed by a set of countable index, say i, and we can described such a vector in 1D with N numbers such as $f_i, i = 1,\ldots,N$ explicitly. This is shown in Figure 36.5(a).

A function $f(x)$, however, can be thought of as being indexed by x in the 1D case. But the index in this case is a continuum, and countably infinite. Thus, it corresponds to a vector of infinite dimension and it lives in an infinite dimensional space.[9]

To make such functions economical in storage, for instance, in 1D example case, we replace the function $f(x)$ by its sampled values at N locations, such that $f(x_i), \quad i = 1,\ldots,N$. Then the values of the function in between the stored points $f(x_i)$ can be obtained by interpolation.[10] Therefore, a function vector $f(x)$, even though it is infinite dimensional, can be approximated by a finite length vector, $\mathbf{f}$. This concept is illustrated in Figure 36.5(b) and (c). This concept can be generalized to a function of 3D space $f(\mathbf{r})$. If $\mathbf{r}$ is sampled over a 3D volume, it can provide an index to a vector $f_i = f(\mathbf{r}_i)$, and thus, $f(\mathbf{r})$ can be thought of as a vector as well.

---

[8] In this course, we have used a boldface letter to denote a 3-vector (viz., vector in 3D space.) But in the rest of the course, a boldface letter can represent an N-dimensional vector.

[9] When these functions are square integrable implying finite "energy", these infinite dimensional spaces are called Hilbert spaces.

[10] This is in fact how special functions like $\sin(x)$, $\cos(x)$, $\exp(x)$, $J_n(x)$, $N_n(x)$, etc, are computed and stored in modern computers. Furthermore, it can be proved that a smooth (or bandlimited) function can be interpolated to exponential accuracy.

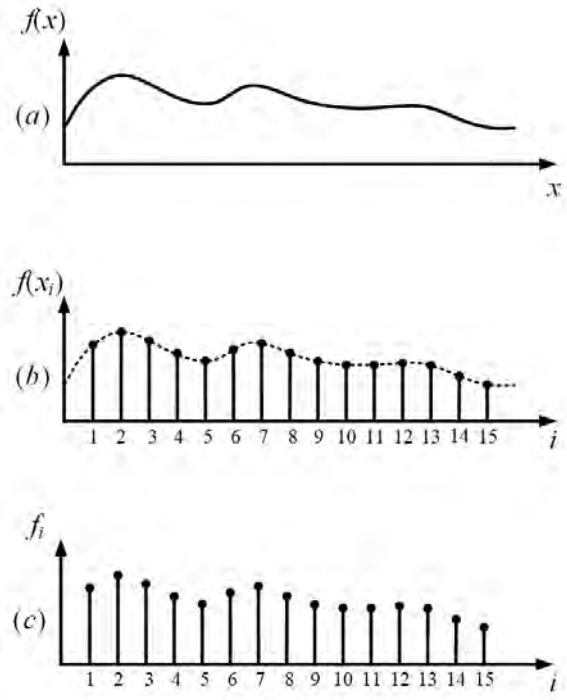


Figure 36.5: A function can be thought of as a vector. (a) A continuum function $f(x)$ plotted as a function of x. (b) A digitally sampled values of the same function. (c) When stored in a computer, it will be stored as an array vector.

36.5 Operator as a Map

36.5.1 Domain and Range Spaces

An operator like $\mathscr{L}$ above can be thought of as a map or a transformation. In this lecture, we will consider linear operators only, and hence, they are like linear matrix operators. A linear operator maps a function f defined in a Hilbert space V to g, another function defined in another Hilbert space W. Mathematically, this is written as

$$\mathscr{L} : V \to W \tag{36.5.1}$$

indicating that $\mathscr{L}$ is a map (or operator) of vectors in the space V to vectors in the space W. Here, V is also called the *domain space* (or domain) of $\mathscr{L}$ while W is the *range space* (or range) of $\mathscr{L}$. We need to familiarize ourselves with this language as research becomes more interdisciplinary.

36.6 Approximating Operator Equations with Matrix Equations

36.6.1 Subspace Projection Methods

One main task of a numerical method is first to approximate an operator equation $\mathscr{L}f = g$ by a matrix equation $\overline{\mathbf{L}} \cdot \mathbf{f} = \mathbf{g}$. To achieve the above, we first let

$$f \cong \sum_{n=1}^{N} a_n f_n \tag{36.6.1}$$

In the above, $f_n, n, \ldots, N$ are known functions called basis functions (or expansion functions analogous to Fourier harmonics in a Fourier series). Now, a_n's are the new unknowns to be sought. Also the above is an approximation, and the accuracy of the approximation depends very much on the original function f. A set of very popular basis functions are functions that form a piece-wise linear interpolation of the function from its nodes. These basis functions are shown in Figure 36.6 in 1D and 2D.

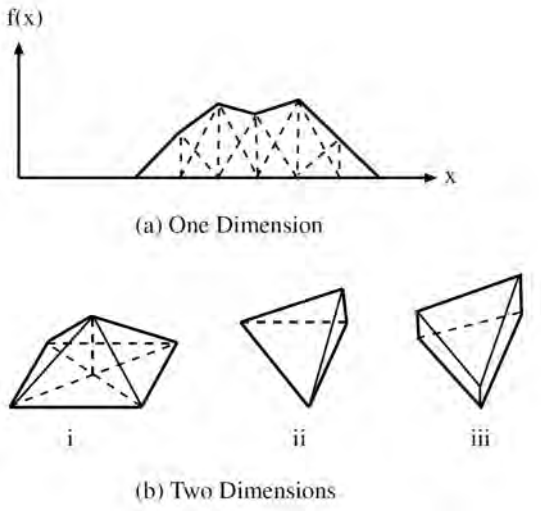


Figure 36.6: Examples of basis function in (a) one dimension, (b) two dimension. Each of these functions are define over a finite domain. Hence, they are also called sub-domain basis functions. They can be thought of as interpolatory functions where the values in between the nodes are obtained by interpolation of the nodal values. When generalized to 3D, these basis functions become tetrahedrons which are harder to draw and visualize.

Upon substituting (36.6.1) into (36.4.1), we obtain

$$\sum_{n=1}^{N} a_n \mathscr{L} f_n = g \tag{36.6.2}$$

Then, upon multiplying (36.6.2) by w_m and integrating over the space that $w_m(\mathbf{r})$ is defined, then we have

$$\sum_{n=1}^{N} a_n \langle w_m, \mathscr{L} f_n \rangle = \langle w_m, g \rangle, m = 1, \ldots, N \tag{36.6.3}$$

In the above, the inner product is defined as

$$\langle f_1, f_2 \rangle = \int d\mathbf{r} f_1(\mathbf{r}) f_2(\mathbf{r}) \tag{36.6.4}$$

where the integration is over the support of the functions, or the space over which the functions are defined.[11] For PDEs these functions are defined over a 2D or 3D coordinate space, while in SIEs, these functions are defined over a surface or a 2D manifold (or a 1D manifold where they become curved lines).[12] In 1D problems, these functions are defined over a 1D coordinate space.

36.6.2 Dual Spaces

The functions $w_m, m = 1, \ldots, N$ in (36.6.3) is known as the weighting functions or testing functions. The testing functions should be chosen so that they can approximate well a function that lives in the range space W of the operator $\mathscr{L}$. Such set of testing functions lives in the *dual space* of the range space. For example, if f_r lives in the range space of the operator $\mathscr{L}$, the set of function f_d, such that the inner product $\langle f_d, f_r \rangle$ exists, forms the dual space of W. If the inner product $\langle f_d, f_r \rangle$ is of infinite value, then f_d is outside the dual space of W.

36.6.3 Matrix and Vector Representations

The above equation (36.6.3) is a matrix equation of the form

$$\overline{\mathbf{L}} \cdot \mathbf{a} = \mathbf{g} \tag{36.6.5}$$

where

$$\begin{aligned} \left[\overline{\mathbf{L}}\right]_{mn} &= \langle w_m, \mathscr{L} f_n \rangle \\ [\mathbf{a}]_n = a_n, \qquad [\mathbf{g}]_m &= \langle w_m, g \rangle \end{aligned} \tag{36.6.6}$$

What has effectively happened here is that given an operator $\mathscr{L}$ that maps a function that lives in an infinite dimensional Hilbert space V, to another function that lives in another infinite dimensional Hilbert space W, via the operator equation $\mathscr{L} f = g$, we have approximated the Hilbert spaces with finite dimensional spaces (subspaces), and finally, obtain a finite dimensional matrix equation that is the representation of the original infinite dimensional operator equation. This is the spirit of the subspace projection method.

In the above, $\overline{\mathbf{L}}$ is the matrix representation of the original operator $\mathscr{L}$ in the subspaces, and $\mathbf{a}$ and $\mathbf{g}$ are the vector representations of f and g, respectively, in their respective subspaces. In this case, it is common to call $L_{mn} = \langle w_m, \mathscr{L} f_n \rangle$ the matrix element of the operator $\mathscr{L}$.

[11] This is known as the reaction inner product [53, 37, 270]. As oppose to most math and physics literature, the energy inner product is used [270] where $\langle f_1, f_2 \rangle = \int d\mathbf{r} f_1^*(\mathbf{r}) f_2(\mathbf{r})$.

[12] A 2D manifold is a curved surface where locally, at a given point, it can be approximated by a flat 2D Euclidean space and similarly for 1D manifold.

When such a method is applied to integral equations, it is usually called the method of moments (MOM). (Surface integral equations are also called boundary integral equations (BIEs) in other fields [269].) When finite discrete basis are used to represent the surface unknowns, it is also called the boundary element method (BEM) [271]. But when this method is applied to solve PDEs, it is called the finite element method (FEM) [272, 273, 274, 193], which is a rather popular method due to its simplicity.

36.6.4 Mesh Generation and Geometry Modeling

To solve for the scattering solution from an arbitrary geometry, we first have to approximate the arbitrary geometry by a mathematical model: this is usually a geometry describable by meshes, which can be mathematically defined as well. In order to approximate a geometry, we need to consider geometries that can be considered as union of line segments, triangle patches, or tetrahedrons. A complicated geometry can be formed from the union of three geometry types. This is called the meshing process (also known as tessellation or discretization).

With these mathematically defined geometries, functions can be defined on an arbitrarily shaped lines, surface, or volume by a finite sum of basis functions, . The basic elements from which more complicated geometries can be built are called simplices.

In 1D, this basic element is a line segment. Union of line segments can be used to approximate arbitrary lines or curves if the curve is segmented fine enough. In 2D, this basic element is a triangle. Union of triangles, with their surface normals pointing in different directions, can be used to model or approximate a surface geometry (see Figure 36.7). In 3D, this basic element is a tetrahedron. Union of tetrahedrons can be use model a 3D volumetric geometry (see Figure 36.8). The smaller these basic elements are, the more accurate the geometry model is. Accuracy of the geometry modeling can be improved by using curvilinear elements as well.

Such meshes are used not only in CEM, but in other fields such as solid mechanics, fluids, and physics. Hence, there are many "solid modeling" commercial software available to generate sophisticated meshes.

Then basis functions, which are defined on these simplices, are used in (36.6.1) and defined to interpolate the field between nodal values in a line segment, triangle, or a tetrahedron.[13]

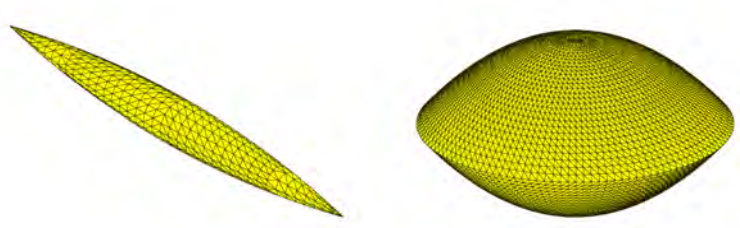


Figure 36.7: An arbitrary surface (also called a 2D manifold) can be meshed by a union of triangles.

[13]Sometimes, the values of the basis functions are defined on edges, e.g., of a triangle, or a tetrahedron [275, 276].

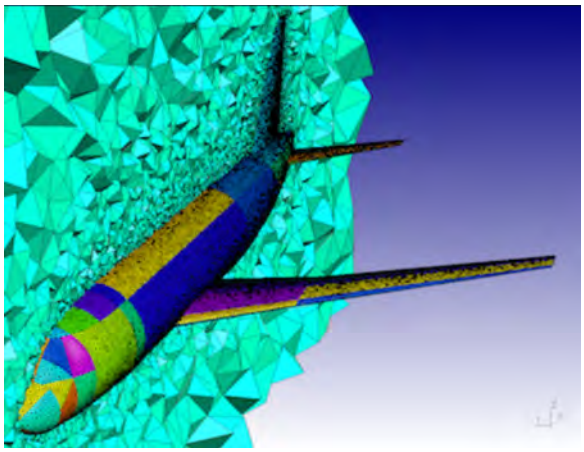


Figure 36.8: A volume region can be meshed by a union of tetrahedra. But the surface of the aircraft is meshed with a union of trianlges (courtesy of gmsh.info).

36.6.5 Differential Equation Solvers versus Integral Equation Solvers

As have been shown, the two classes of numerical solvers for Maxwell's equations are differential equation solvers and integral equation solvers. Differential equation solvers are generally easier to implement. As shall be shown in the next lecture, they can also be easily implemented using finite difference solver. The unknowns in a differential equation solver are the fields. Since the fields permeate all of space, the unknowns are volumetrically distributed. When the fields are discretized by representing them by their point values in space, they require a large number of unknowns to represent. The plus side is that the matrix system associated with a differential equation solver is usually sparse, requiring less storage and less time to solve.

As has been shown, integral equation solvers are formulated using Green's functions. In other words, integral equations are derived from Maxwell's equations using Green's function (or dyadic Green's function), where the unknowns now are surface unknowns such as surface electric and magnetic currents, or volume uknowns. Therefore, the unknowns are generally smaller in number, living only on the surface of a scatterer (or they occupy a smaller part of space). Hence, they can be approximated by a smaller set of unknowns. Consequently, the matrix systems generally are smaller. Once the currents are found, then the fields they generate can also be computed.[14]

Since the derivation of integral equations requires the use of Green's functions, they are in general singular when $\mathbf{r} = \mathbf{r}'$, or when the observation point (observation point) $\mathbf{r}$ and the source point $\mathbf{r}'$ coincide. Care has to be taken to discretize the integral equations in the neighborhood of the singularity. Moreover, a Green's function connects every current source point on the surface of a scatterer with every other source points yielding a dense matrix system. But fast methods have been developed to solve such dense matrix systems [9].

[14] Volume integral equations are also used, but the reduction in unknown count is not as dramatic.

36.7 Matrix Solution by Matrix-Free Method

36.7.1 Computational Complexity and Curse of Dimensionality

When computers are used to solve real-world problems, the number of unknowns needed are in general humongous. It is easy to see that problems in 1D can be modeled by a small number of unknowns. In CEM, the yardstick is usually wavelengths. If n unknowns are used to approximate a function in a wavelength, it is clear that the unknown count is n^d where d is the dimensionality of the problem. In three dimensions, the unknown count scales as n^3 (or $O(n^3)$, read 'of order n^3')), which is a terrible scaling when n is large and d is not small. This is also known as the curse of dimensionality.

It is seen that the solution of Maxwell's equations can be reduced to solving a matrix equation $\overline{\mathbf{L}} \cdot \mathbf{a} = \mathbf{g}$, where $\overline{\mathbf{L}}$ is a $N \times N$ matrix. The old way of solving a matrix system is to use Gauss elimination, or Cramer's rule. Unfortunately, the number of arithmetics operations needed to solve the matrix system using these old rules are proportional of N^3 (or of $O(N^3)$) This is again, a terrible scaling, exhausting easily our computational resources.

The modern way of solving a matrix system is to find one where the CPU scaling scales as $O(N^\alpha)$ where α is made as close to one as possible. For an algorithm, the rate at which the CPU time grows as N is known as the computational complexity of the algorithm. For instance, we say that Gaussian elimination has a computational complexity of $O(N^3)$ which is a terrible complexity.

Another issue that designers of modern day computational algorithms have to grapple with is the issue of memory complexity. For instance, in forming a matrix equation, one has to form (or fill) a matrix system. The matrix and vector elements are generated in accordance to the formulas for their representations as shown in (36.6.5) and (36.6.6). Since there are N^2 elements in a matrix system, filling or forming the matrix system will take $O(N^2)$ time and memory usage. This is a terrible scaling especially when compounded by the curse of dimensionality. Three D problems especially are notorious hogger of computer resources in both CPU and memory usage!

36.7.2 Matrix Solutions

As aforementioned, given a matrix equation, there are many ways to seek its solution. The simplest way is to find the inverse of the matrix operator by direct inversions (e.g., using Gaussian elimination [181] or Kramer's rule [277]). But on the down side, they have computational complexity[15] of $O(N^3)$, and requiring storage of $O(N^2)$. Due to the poor computational and memory complexity of direct inversion, when N is large, other methods have to be sought. One way is to seek a matrix-free method! In a matrix-free method, the matrix system is never generated, but an algorithm to produce the matrix-vector product $\overline{\mathbf{L}} \cdot \mathbf{a} = \mathbf{c}'$ is needed. The memory needed to store $\mathbf{c}'$ is $O(N)$ which is much less than that for storing the matrix $\overline{\mathbf{L}}$.

To this end, it is better to convert the solving of a matrix equation into an optimization problem which is matrix-free. These methods can be designed so that a much larger system can be solved with an existing resource of a digital computer. Optimization problem results in finding the stationary point of a functional.[16] First, we will figure out how to find such a functional.

[15]The scaling of computer time with respect to the number of unknowns (degrees of freedom) is known in the computer parlance as computational complexity.

[16]Functional is usually defined as a function of a function [52, 37]. Here, we include a function of a vector to be

Consider a matrix equation given by

$$\overline{\mathbf{L}} \cdot \mathbf{f} = \mathbf{g} \tag{36.7.1}$$

For simplicity, we consider $\overline{\mathbf{L}}$ as a symmetric matrix.[17] Then the corresponding functional or cost function is

$$I = \mathbf{f}^t \cdot \overline{\mathbf{L}} \cdot \mathbf{f} - 2\mathbf{f}^t \cdot \mathbf{g} \tag{36.7.2}$$

Such a functional is a quadratic functional because it is analogous to $I = Lx^2 - 2xg$, which is quadratic, in its simplest 1D rendition.

To find its optimal value or its stationary point, we take the first variation with respect to $\mathbf{f}$, namely, we let $\mathbf{f} = \mathbf{f}_o + \delta\mathbf{f}$. Then we substitute this into the above, and collect the leading order and first order terms. As a result, we find the first order approximation of the functional I as

$$\delta I = \delta\mathbf{f}^t \cdot \overline{\mathbf{L}} \cdot \mathbf{f}_o + \mathbf{f}_o^t \cdot \overline{\mathbf{L}} \cdot \delta\mathbf{f} - 2\delta\mathbf{f}^t \cdot \mathbf{g} \tag{36.7.3}$$

If $\overline{\mathbf{L}}$ is a symmetric matrix, the first two terms are the same, which is easily verified by taking the transpose of one of them, and using that the transpose of a scalar is itself. Then the above just becomes

$$\delta I = 2\delta\mathbf{f}^t \cdot \overline{\mathbf{L}} \cdot \mathbf{f}_o - 2\delta\mathbf{f}^t \cdot \mathbf{g} \tag{36.7.4}$$

For $\mathbf{f}_o$ to be the optimal point or the stationary point, then its first variation has to be zero, or that $\delta I = 0$. Since $\delta\mathbf{f}^t$ above is arbitrary, as a result, we conclude that at the optimal point (or the stationary point), the following equation has to be satisfied:

$$\overline{\mathbf{L}} \cdot \mathbf{f}_o = \mathbf{g} \tag{36.7.5}$$

Hence, the optimal point to the quadratic functional I in (36.7.2) is the solution to (36.7.1) or (36.7.5).

36.7.3 Gradient of a Functional

The above method, when applied to an infinite dimensional Hilbert space problem, is called the variational method, but the main ideas are similar. The wonderful idea about such a method is that instead of doing direct inversion of a matrix system (which is expensive), one can search for the optimal point or stationary point of the quadratic functional using gradient search or gradient descent methods or some optimization method.

It turns out that the gradient of a quadratic functional can be found quite easily. Also it is cheaper to compute the gradient of a functional than to find the inverse of a matrix operator. To do this, it is better to write out functional using index (or indicial, or Einstein) notation [278]. In this notation, summations over repeated indices are implied. Then, the functional first variation δI in (36.7.4) becomes

$$\delta I = 2\delta f_j L_{ji} f_i - 2\delta f_j g_j \tag{36.7.6}$$

a functional as well.

[17] Functional for the asymmetric case can be found in *Chew, Waves and Fields in Inhomogeneous Media*, Chapter 5 [37].

Also as aforementioned, in this notation, the summation symbol is dropped, and summations over repeated indices are implied. In the above, we neglect to distinguish between $\mathbf{f}_o$ and $\mathbf{f}$. It is implied that $\mathbf{f}$ represents the optimal point. In this manner, it is easier to see what a functional derivative is. We can differentiate the above with respect to f_j easily to arrive at

$$\frac{\partial I}{\partial f_j} = 2L_{ij}f_i - 2g_j \tag{36.7.7}$$

Notice that the remaining equation has one index j remaining in index notation, meaning that it is a vector equation. We can reconstitute the above using our more familiar matrix notation that the above is similar to

$$\frac{\delta I}{\delta \mathbf{f}} = \nabla_{\mathbf{f}} I = 2\overline{\mathbf{L}} \cdot \mathbf{f} - 2\mathbf{g} \tag{36.7.8}$$

The left-hand side is a notation for the functional derivative or the gradient of a functional in a multi-dimensional space which is a vector obviated by indicial notation. And the right-hand side is the expression for calculating this gradient. One needs only to perform a matrix-vector product to find this gradient. The cost of a matrix-vector product is at most $O(N^2)$ for dense matrices, and can be as low as $O(N)$ for sparse matrices. Hence, the computational complexity of finding this gradient is $O(N^2)$ at worst if $\overline{\mathbf{L}}$ is a dense matrix, and as low as $O(N)$ if $\overline{\mathbf{L}}$ is a sparse matrix.[18] In a gradient search method, such a gradient is calculated repeatedly until the optimal point is found. Such methods are called iterative methods.

If the optimal point can be found in N_{iter} iterations, then the CPU time scales as $N_{\text{iter}}N^\alpha$ where $1 < \alpha < 2$. There is a clever gradient search algorithm, called the *conjugate gradient method* that can find the exact optimal point in $N_{\text{iter}} = N$ in exact arithmetics. But exact solution is not needed in an optimal solution: an approximate solution suffices. In many gradient search method, it suffices obtain an approximate solution where the error is acceptable after N_{iter} where $N_{\text{iter}} \ll N$ is possible. Thus the total solution time (or solve time) is $N_{\text{iter}}N^\alpha \ll NN^\alpha \ll N^3$ is possible, resulting in great savings in computer time, especially if $\alpha = 1$. This is the case for FEM [193], [279], [280], [281], [274], and fast multipole algorithm [282], [283].[19]

What is more important is that this method does not require the storage of the matrix $\overline{\mathbf{L}}$, but a computer code that produces the vector $\mathbf{g}_o = \overline{\mathbf{L}} \cdot \mathbf{f}$ as an output, with $\mathbf{f}$ as an input. Both $\mathbf{f}$ and $\mathbf{g}_o$ require only $O(N)$ memory storage. Such methods are called matrix-free methods. Even when $\overline{\mathbf{L}}$ is a dense matrix, which is the case if it is the matrix representation of matrix representation of some Green's function, fast methods now exist to perform the dense matrix-vector product in $O(N \log N)$ operations.[20]

The value I is also called the cost function, and its minimum is sought in finding the solution by gradient search methods. Detail discussions of these methods are given in [284]. Figure 36.9 shows the contour plot of a cost function in 2D. When the condition number[21] of the matrix $\overline{\mathbf{L}}$ is large (implying that the matrix is ill-conditioned), the contour plot resembles a deep and narrorw valley. And hence, the gradient search method will tend to zig-zag along the way as it finds the

[18] This is the case for many differential equation solvers such as finite-element method or finite-difference method.

[19] It is to be noted that many fast algorithms have computational complexity of $O(N \log N)$, such as FFT and fast multipole algorithm. It is important to note that $N \log N < N^{1+\alpha}$, where $\alpha > 0$ as $N \to \infty$.

[20] Chew et al, *Fast and Efficient Algorithms in CEM* [9].

[21] This is the ratio of the largest eigenvalue of the matrix to its smallest eigenvalue.

optimal solution. This implies that convergence is slow for matrices with large condition numbers. It is seen graphically that when the condition number of a matrix system is large, it is deleterious to the convergence of gradient search methods [285, 286].

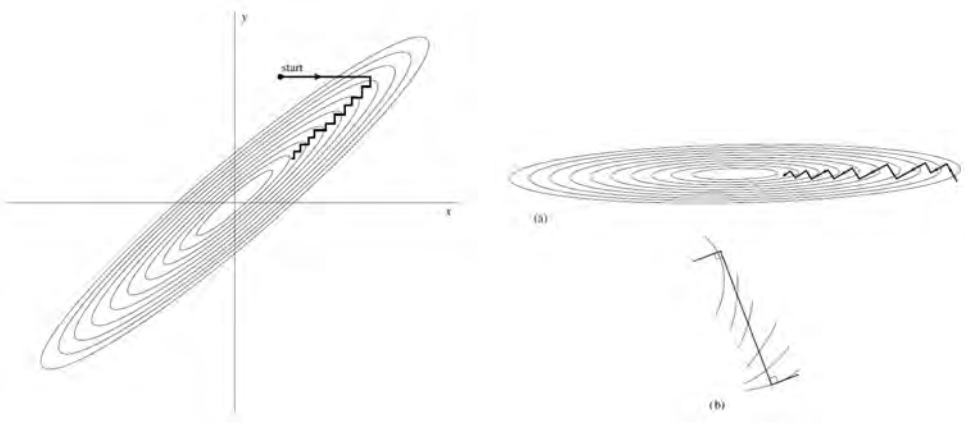


Figure 36.9: Plot of a 2D cost function, $I(x, y)$ for an ill-conditioned system (courtesy of Numerical Recipe [284]). A higher dimensional plot of this cost function will be difficult. The zig-zag search path is the hall-mark of an ill-conditioned system. This cost function has only two eigenvalues, one of which is larger than the other one.

Figure 36.10 shows a cartoon picture in 2D of the histories of different search paths from a machine-learning example where a cost functional similar to I has to be minimized. Finding the optimal point or the minimum point of a general functional is still a hot topic of research: it is important in artificial intelligence as well as in solving large system of linear algebraic equations.

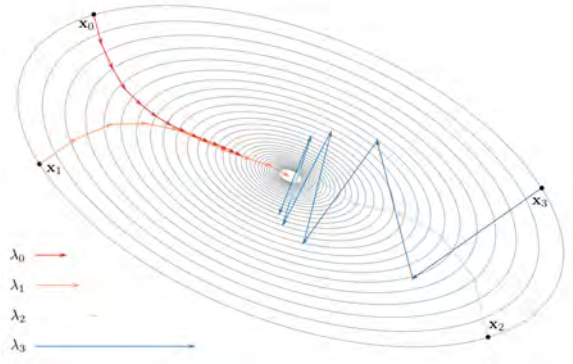


Figure 36.10: Gradient search or gradient descent method is finding an optimal point. As is obvious, the initial value is instrumental in converging into the correct solution. (courtesy of Y. Ioannou: https://blog.yani.io/sgd/).

Exercises for Lecture 36

Problem 36-1: This exercise refers to Chapter 36 of the lecture notes.

(i) Derive eq. (36.2.2) from Maxwell's equations.

(ii) In the lecture notes, we assume that $\overline{\mathbf{L}}$ is a symmetric matrix. But now, assume that the operator $\overline{\mathbf{L}}$ in (36.7.1) is not symmetric. The optimal solution is obtained by solving two equations. Construct a functional I such that is optimal point is the solution to eq. (36.7.1) and another auxiliary equation. Show that such a functional can be defined as

$$I = \mathbf{w}^t \cdot \overline{\mathbf{L}} \cdot \mathbf{f} - \mathbf{w}^t \cdot \mathbf{g} - \mathbf{g}_a^t \cdot \mathbf{f}$$

By taking the first variation of the above functional with respect to f and w, show that the optimal solutions that minimize the above functional are solutions to the equations

$$\overline{\mathbf{L}} \cdot \mathbf{f} = \mathbf{g}$$

$$\overline{\mathbf{L}}^t \cdot \mathbf{w} = \mathbf{g}_a$$

(iii) Using index notation, for a symmetric system discussed in the lecture notes, show that the gradient of a functional I in the N dimensional space is given by (36.7.8).

Problem 36-2: This exercise refers to Chapter 36 of the lecture notes.

(i) Derive eq. (36.3.5) and explain why the second term on the right-hand side corresponds to the scattered field.

(ii) Explain why the scattered field satisfies the radiation condition.

Chapter 37

Finite Difference Method, Yee Algorithm

In this lecture, we will introduce one of the simplest methods to solve Maxwell's equations numerically. This is the finite-difference time-domain method first proposed by Yee [287] and popularized by Taflove [288]. Because of its simplicity, a simple Maxwell's equations solver can be coded in one afternoon. Thus almost every physics or electrical engineering laboratory has a home-grown version of the finite-difference time-domain solver. This method is the epitome of that "simplicity rules."[1] Professor Hermann Haus at MIT used to say: find the simplest method to do things. Complicated methods will be forgotten, but the simplest ones will prevail. This is also reminiscent of Einstein's saying, "Everything should be made as simple as possible, but no simpler!"

37.1 Finite-Difference Time-Domain Method

To obtain the transient (or time-domain) solution of the wave equation for a more general, inhomogeneous medium, a numerical method has to be used. The finite-difference time-domain (FDTD) method, a numerical method, is particularly suitable for solving transient problems. Compounded by rapid growth in computer speed, with its versatility, it has been used with great success in solving many practical problems. This method is based on the simple Yee algorithm [287] and has been vastly popularized by Taflove [288, 289].

In the finite-difference method, continuous space-time is replaced with a discrete space-time. Thus in the discrete space-time, partial differential equations are replaced with finite difference equations. These finite difference equations are readily implemented on a digital computer. Furthermore, an iterative or time-stepping scheme can be implemented without having to store or solve large matrices, resulting in great savings in computer time and memory. In addition, the matrix for the system of equations is never generated making this a matrix-free method: There is no need to store the matrix system for matrix and memory management as one writes this

[1] "rule" is used as a verb.

numerical solver. More recently, the development of parallel processor architectures in computers has also further enhanced the efficiency of the finite-difference time-domain scheme [290].

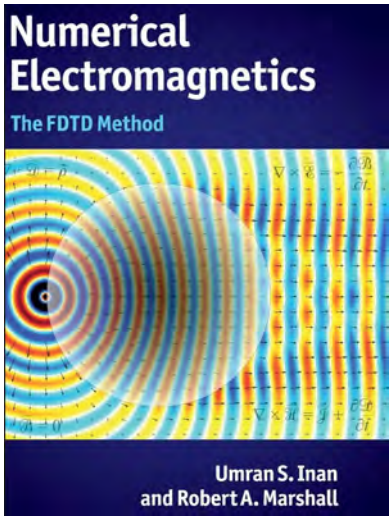


Figure 37.1: The finite-difference time-domain method is so popular, and so simple that many authors would choose to adorn their book covers with the beautiful graphics of FDTD simulations, as shown in the fiure here.

The finite-difference method is also described in numerous early works (see, for example, Potter 1973 [291]; Taflove 1988 [288]; Ames 2014 [292]; and Morton 2019 [293].

37.1.1 The Finite-Difference Approximation

Consider first a scalar wave equation of the form

$$\frac{1}{c^2(\mathbf{r})}\frac{\partial^2}{\partial t^2}\phi(\mathbf{r},t) = \mu(\mathbf{r})\nabla\cdot\mu^{-1}(\mathbf{r})\nabla\phi(\mathbf{r},t). \tag{37.1.1}$$

The above equation appears in scalar acoustic waves or a 2D electromagnetic waves in inhomogeneous media [152, 37]. It clearly has no closed-form solution.

To convert the above into a form that can be solved by a digital computer easily, first, one needs to find finite-difference approximations to the time derivatives. The time derivative can be approximated in many ways. For example, a derivative can be approximated by forward, backward, and central finite difference formulas [294] (see Figure 37.2).

$$\text{Forward difference:}\quad \frac{\partial\phi(\mathbf{r},t)}{\partial t} \approx \frac{\phi(\mathbf{r},t+\Delta t)-\phi(\mathbf{r},t)}{\Delta t}, \tag{37.1.2}$$

$$\text{Backward difference:}\quad \frac{\partial\phi(\mathbf{r},t)}{\partial t} \approx \frac{\phi(\mathbf{r},t)-\phi(\mathbf{r},t-\Delta t)}{\Delta t}, \tag{37.1.3}$$

$$\text{Central difference:}\quad \frac{\partial\phi(\mathbf{r},t)}{\partial t} \approx \frac{\phi(\mathbf{r},t+\frac{\Delta t}{2})-\phi(\mathbf{r},t-\frac{\Delta t}{2})}{\Delta t}, \tag{37.1.4}$$

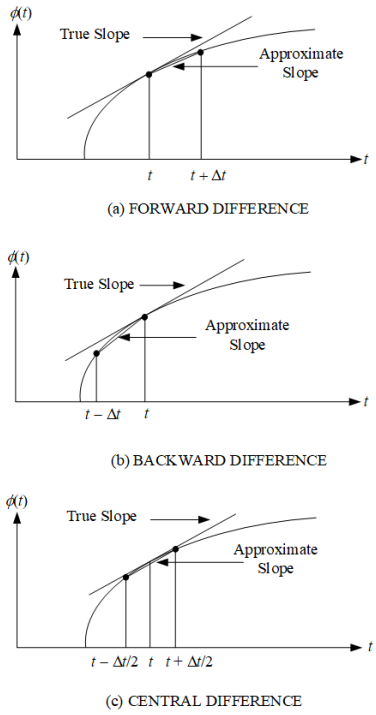


Figure 37.2: Different finite-difference approximations for the time derivative. One can eye-ball the above plots and see that the central difference formula is the best. This can be further confirmed by a Taylor series analysis.

where Δt is a small number. Of the three methods of approximating the time derivative, the central-difference scheme is the best approximation, as is evident from Figure 37.2. The errors in the forward and backward differences are $O(\Delta t)$ (or first-order error) while the central-difference approximation has an error $O[(\Delta t)^2]$ (or second-order error). This can be easily verified by Taylor-series expanding the right-hand sides of (37.1.2) to (37.1.4).

Consequently, using the central-difference formula twice, we arrive at the approximation for the second derivative as

$$\frac{\partial^2}{\partial t^2}\phi(\mathbf{r},t) \approx \frac{\partial}{\partial t}\left[\frac{\phi(\mathbf{r},t+\frac{\Delta t}{2}) - \phi(\mathbf{r},t-\frac{\Delta t}{2})}{\Delta t}\right] \tag{37.1.5}$$

$$\approx \frac{\phi(\mathbf{r},t+\Delta t) - 2\phi(\mathbf{r},t) + \phi(\mathbf{r},t-\Delta t)}{(\Delta t)^2}. \tag{37.1.6}$$

Next, if the function $\phi(\mathbf{r},t)$ is indexed on discrete time steps on the t axis, such that for $t = l\Delta t$, then $\phi(\mathbf{r},t) = \phi(\mathbf{r},l\Delta t) = \phi^l(\mathbf{r})$. Here, l is an integer used to count the time steps. Using this

notation, Equation (37.1.6) then becomes

$$\frac{\partial^2}{\partial t^2}\phi(\mathbf{r},t) \approx \frac{\phi^{l+1}(\mathbf{r}) - 2\phi^l(\mathbf{r}) + \phi^{l-1}(\mathbf{r})}{(\Delta t)^2}. \tag{37.1.7}$$

37.1.2 Time Stepping or Time Marching

With this notation and approximations, (37.1.1) can be approximated by a time-stepping (or time-marching) formula, namely,

$$\phi^{l+1}(\mathbf{r}) \cong c^2(\mathbf{r})(\Delta t)^2\mu(\mathbf{r})\nabla\cdot\mu^{-1}(\mathbf{r})\nabla\phi^l(\mathbf{r}) + 2\phi^l(\mathbf{r}) - \phi^{l-1}(\mathbf{r}). \tag{37.1.8}$$

Therefore, given the knowledge of $\phi(\mathbf{r},t)$ at $t = l\Delta t$ or $\phi^l(\mathbf{r})$, and $t = (l-1)\Delta t$, or $\phi^{l-1}(\mathbf{r})$ for all $\mathbf{r}$, one can deduce $\phi(\mathbf{r},t)$ at $t = (l+1)\Delta t$, or $\phi^{l+1}(\mathbf{r})$ for all $\mathbf{r}$. In other words, given the initial values of $\phi(\mathbf{r},t)$ at, for example, $t = 0$ and $t = \Delta t$, $\phi(\mathbf{r},t)$ can be deduced for all subsequent times, provided that the time-stepping formula is stable.

At this point, the right-hand side of (37.1.8) involves the space derivatives. There exist a plethora of ways to approximate and calculate the right-hand side of (37.1.8) numerically. Here, we shall illustrate again the use of the finite-difference method to calculate the right-hand side of (37.1.8). Before proceeding further, note that the space derivatives on the right-hand side in cartesian coordinates are

$$\mu(\mathbf{r})\nabla\cdot\mu^{-1}(\mathbf{r})\nabla\phi(\mathbf{r}) = \mu\frac{\partial}{\partial x}\mu^{-1}\frac{\partial}{\partial x}\phi + \mu\frac{\partial}{\partial y}\mu^{-1}\frac{\partial}{\partial y}\phi + \mu\frac{\partial}{\partial z}\mu^{-1}\frac{\partial}{\partial z}\phi. \tag{37.1.9}$$

Then, one can approximate, using central differencing that

$$\frac{\partial}{\partial z}\phi(x,y,z) \approx \frac{1}{\Delta z}\left[\phi\left(x,y,z+\frac{\Delta z}{2}\right) - \phi\left(x,y,z-\frac{\Delta z}{2}\right)\right], \tag{37.1.10}$$

Consequently, using central differencing two times,

$$\begin{aligned}\frac{\partial}{\partial z}\mu^{-1}\frac{\partial}{\partial z}\phi(x,y,z) \approx \frac{1}{(\Delta z)^2}\Bigg\{&\mu^{-1}\left(z+\frac{\Delta z}{2}\right)\phi(x,y,z+\Delta z)\\ &-\left[\mu^{-1}\left(z+\frac{\Delta z}{2}\right)+\mu^{-1}\left(z-\frac{\Delta z}{2}\right)\right]\phi(x,y,z)\\ &+\mu^{-1}\left(z-\frac{\Delta z}{2}\right)\phi(x,y,z-\Delta z)\Bigg\}.\end{aligned} \tag{37.1.11}$$

Furthermore, after denoting $\phi(x,y,z) = \phi_{m,n,p}$, $\mu(x,y,z) = \mu_{m,n,p}$, on a discretized grid point at $x = m\Delta x$, $y = n\Delta y$, $z = p\Delta z$, we have $(x,y,z) = (m\Delta x, n\Delta y, p\Delta z)$. Then in finite-difference notations,

$$\begin{aligned}\frac{\partial}{\partial z}\mu^{-1}\frac{\partial}{\partial z}\phi(x,y,z) \approx \frac{1}{(\Delta z)^2}\Big[&\mu^{-1}_{m,n,p+\frac{1}{2}}\phi_{m,n,p+1}\\ &-\left(\mu^{-1}_{m,n,p+\frac{1}{2}}+\mu^{-1}_{m,n,p-\frac{1}{2}}\right)\phi_{m,n,p}+\mu^{-1}_{m,n,p-\frac{1}{2}}\phi_{m,n,p-1}\Big].\end{aligned} \tag{37.1.12}$$

This cumbersome and laborious looking equation can be abbreviated if we define a central difference operator as[2]

$$\bar{\partial}_z \phi_{m,n,p} = \frac{1}{\Delta z}\left(\phi_{m,n,p+\frac{1}{2}} - \phi_{m,n,p-\frac{1}{2}}\right) \tag{37.1.13}$$

Then the right-hand side of the (37.1.12) can be written succinctly as

$$\frac{\partial}{\partial z}\mu^{-1}\frac{\partial}{\partial z}\phi(x,y,z) \approx \bar{\partial}_z \mu^{-1}_{m,n,p}\bar{\partial}_z \phi_{m,n,p} \tag{37.1.14}$$

With similar approximations to the other terms in (37.1.9), (37.1.8) is now compactly written as

$$\begin{aligned}\phi^{l+1}_{m,n,p} = (\Delta t)^2 c^2_{m,n,p}\mu_{m,n,p}\left[\bar{\partial}_x \mu^{-1}_{m,n,p}\bar{\partial}_x + \bar{\partial}_y \mu^{-1}_{m,n,p}\bar{\partial}_y + \bar{\partial}_z \mu^{-1}_{m,n,p}\bar{\partial}_z\right]\phi^l_{m,n,p} \\ + 2\phi^l_{m,n,p} - \phi^{l-1}_{m,n,p}.\end{aligned} \tag{37.1.15}$$

The above can be readily implemented on a computer for time stepping. Notice however, that the use of central differencing results in the evaluation of medium property μ at half grid points. This is inconvenient, as the introduction of material values at half grid points increases computer memory used. Hence, it is customary to store the medium values at the integer grid points for ease of book-keeping, and to deduce the values at half-grid points using the following approximations

$$\mu_{m+\frac{1}{2},n,p} \simeq \frac{1}{2}(\mu_{m+1,n,p} + \mu_{m,n,p}), \tag{37.1.16}$$

$$\mu_{m+\frac{1}{2},n,p} + \mu_{m-\frac{1}{2},n,p} \simeq 2\mu_{m,n,p}, \tag{37.1.17}$$

and so on. Moreover, if μ is a smooth function of space, it is easy to show that the errors in the above approximations are of second order by Taylor series expansions.

For a homogeneous medium, with $\Delta x = \Delta y = \Delta z = \Delta s$, namely, we assume the space steps to be equal in all directions, (37.1.15) written explicitly then becomes

$$\begin{aligned}\phi^{l+1}_{m,n,p} = \left(\frac{\Delta t}{\Delta s}\right)^2 c^2 \left[\phi^l_{m+1,n,p} + \phi^l_{m-1,n,p} + \phi^l_{m,n+1,p} + \phi^l_{m,n-1,p} + \phi^l_{m,n,p+1}\right. \\ \left. + \phi^l_{m,n,p-1} - 6\phi^l_{m,n,p}\right] + 2\phi^l_{m,n,p} - \phi^{l-1}_{m,n,p}.\end{aligned} \tag{37.1.18}$$

Notice then that with the central-difference approximation, the value of $\phi^{l+1}_{m,n,p}$ is dependent only on $\phi^l_{m,n,p}$, and its nearest neighbors, $\phi^l_{m\pm1,n,p}$, $\phi^l_{m,n\pm1,p}$, $\phi^l_{m,n,p\pm1}$, and $\phi^{l-1}_{m,n,p}$, its value at the previous time step. Moreover, in the finite-difference scheme outlined above, no matrix inversion is required at each time step. Such a scheme is also known as an explicit scheme. The use of an explicit scheme is a major advantage of the finite-difference method compared to the finite-element methods. Consequently, in order to update N grid points using (37.1.15) or (37.1.18), $O(N)$ multiplications are required for each time step. In comparison, $O(N^3)$ multiplications are required to invert an $N \times N$ full matrix, e.g., using Gaussian elimination. The simplicity and efficiency of these finite-difference algorithms have made them vastly popular.

[2]This is in the spirit of [295].

37.1.3 Stability Analysis

The implementation of the finite-difference time-domain scheme using time-marching does not always lead to a stable scheme. Hence, in order for the solution to converge, the time-stepping scheme must at least be stable. Consequently, it is useful to find the condition under which this numerical scheme is stable. To do this, one performs the von Neumann stability analysis (von Neumann 1943 [296]) on Equation (37.1.18). We will assume the medium to be homogeneous to simplify the analysis.

As shown in the previous Chapter 35, Section 35.1, a point source gives rise to a spherical wave that can be expanded as sum of plane waves in different directions. It also implies that any wave emerging from sources can be expanded as sum of plane waves. This is the spirit of the spectral expansion method in the previous chapter: if a scheme is not stable for a plane wave, it would not be stable for any wave. Consequently, to perform the stability analysis, we assume a propagating plane wave (or a Fourier mode) as a trial solution. Namely, we let

$$\phi(x,y,z,t) = A_0 e^{ik_x x + ik_y y + ik_z z - i\omega t}, \tag{37.1.19}$$

The above is clearly a solution to the scalar wave equation. In discretized form, we let $x = m\Delta x$, $y = n\Delta y$, $z = p\Delta z$, $t = l\Delta t$ and then letting $\Delta x = \Delta y = \Delta z = \Delta s$, we have a discrete version of the Fourier mode;

$$\phi^l_{m,n,p} = A_0 e^{ik_x m\Delta s + ik_y n\Delta s + ik_z p\Delta s - i\omega l\Delta t}. \tag{37.1.20}$$

It turns out the discrete Fourier modes are eigenfunctions of the finite-difference operators. Using (37.1.20), it is easy to show that for the x space derivative,

$$\begin{aligned}(\Delta s)^2 \bar{\partial}_x^2 \phi^l_{m,n,p} = \phi^l_{m+1,n,p} - 2\phi^l_{m,n,p} + \phi^l_{m-1,n,p} &= 2[\cos(k_x\Delta s) - 1]\phi^l_{m,n,p} \\ &= -4\sin^2\left(\frac{k_x\Delta s}{2}\right)\phi^l_{m,n,p}.\end{aligned} \tag{37.1.21}$$

The above indicates that the space finite difference operator acting on the discrete Fourier mode is just a constant times the Fourier mode. This implies that a Fourier mode is an eigenfunction of the finite difference operator! The space derivatives in y and z directions can be similarly derived.

The second order time derivative in the wave equation can be similarly approximated by a finite-difference equation, and it is equal to

$$\frac{\partial^2}{\partial t^2}\phi(\mathbf{r},t)(\Delta t)^2 \approx \phi^{l+1}_{m,n,p} - 2\phi^l_{m,n,p} + \phi^{l-1}_{m,n,p} = -4\sin^2\left(\frac{\omega\Delta t}{2}\right)\phi^l_{m,n,p} \tag{37.1.22}$$

We need to find the finite-difference approximation of the wave equation, namely that

$$\nabla^2\phi(\mathbf{r},t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\phi(\mathbf{r},t) = 0 \tag{37.1.23}$$

Using (37.1.21) and its equivalence for x, y, z space derivatives, and then (37.1.22) for the time

derivatives, it follows that the finite-difference approximation of the above wave equation gives us

$$4\sin^2\left(\frac{\omega\Delta t}{2}\right)\phi^l_{m,n,p} = 4\left(\frac{\Delta t}{\Delta s}\right)^2 c^2\left[\sin^2\left(\frac{k_x\Delta s}{2}\right)+\sin^2\left(\frac{k_y\Delta s}{2}\right)+\sin^2\left(\frac{k_z\Delta s}{2}\right)\right]\phi^l_{m,n,p}$$
$$= 4r^2s^2\phi^l_{m,n,p}, \tag{37.1.24}$$

where

$$r = \left(\frac{\Delta t}{\Delta s}\right)c, \qquad s^2 = \sin^2\left(\frac{k_x\Delta s}{2}\right)+\sin^2\left(\frac{k_y\Delta s}{2}\right)+\sin^2\left(\frac{k_z\Delta s}{2}\right). \tag{37.1.25}$$

Equation (37.1.24) implies that, for nonzero $\phi^l_{m,n,p}$,[3]

$$\sin^2\left(\frac{\omega\Delta t}{2}\right) = r^2s^2, \tag{37.1.26}$$

The above is the dispersion relation relating ω, the frequency of the Fourier mode with k_x, k_y, k_z the Fourier wave numbers.

In order for the solution to be stable, ω has to be real when we solve the above equation for ω. Hence, it is necessary that the right-hand side of the above equation is less than 1 so that ω is real. This implies that if

$$r^2s^2 < 1, \tag{37.1.27}$$

then stability is ensured. Since from (37.1.25), $s^2 \le 3$ for all k_x, k_y, and k_z, from (37.1.27). Also, from (37.1.27), we conclude that

$$r < \frac{1}{s}$$

But we also know that from the definition of s in (37.1.25) that

$$\frac{1}{s} \ge \frac{1}{\sqrt{3}}$$

In other words, the right-hand side of the above is the lower bound for $1/s$. The above two inequalities will be satisfied if the general condition is

$$r < \frac{1}{\sqrt{3}} < \frac{1}{s}, \qquad \text{or} \qquad \Delta t < \frac{\Delta s}{c\sqrt{3}}. \tag{37.1.28}$$

after using that $r = c\Delta t/(\Delta s)$. The above is the general condition for stability. The above analysis is for 3 dimensional problems. It is clear from the above analysis that for an n-dimensional problem where $n = 1, 2, 3$, then

$$\Delta t < \frac{\Delta s}{c\sqrt{n}}. \tag{37.1.29}$$

[3]For those who are more mathematically inclined, we are solving an eigenvalue problem in disguise. Remember that a function is a vector, even after it has been discretized:)

One may ponder on the physical meaning of this inequality further: but it is only natural that the time step Δt has to be bounded from above. Otherwise, one arrives at the ludicrous notion that the time step can be arbitrarily large thus violating causality.

Moreover, if the grid points of the finite-difference scheme are regarded as a simple cubic lattice, then the distance $\Delta s/\sqrt{n}$ is also the distance between the closest lattice planes through the simple cubic lattice. Notice that the time for the wave to travel between these two lattice planes is $\Delta s/(c\sqrt{n}\,)$. Consequently, the stability criterion (37.1.29) implies that the time step Δt has to be less than the shortest travel time for the wave between the lattice planes in order to satisfy causality. In other words, if the wave is time-stepped ahead of the time on the right-hand side of (37.1.29), instability ensues, because of the violation of causality.

The above is also known as the CFL (Courant, Friedrichs, and Lewy 1928 [297]) stability criterion. It could be easily modified for $\Delta x \neq \Delta y \neq \Delta z$ [289].The above analysis implies that we can pick a larger time step if the space steps are larger. A larger time step will allow one to complete generating a time-domain response rapidly. However, one cannot arbitrary make the space step large due to grid-dispersion error, as shall be discussed next.

37.1.4 Grid-Dispersion Error

When a finite-difference scheme is stable, it still may not be accurate to produce good results due to the errors in the finite-difference approximations. Hence, it is useful to ascertain the errors in terms of the size of the grid and the time step. An easy error to analyze is the *grid-dispersion error*. In a homogeneous, dispersionless medium, all plane waves propagate with the same phase velocity. However, in the finite-difference approximation, all plane waves will not propagate at the same phase velocity due to the grid-dispersion error.

As a consequence, a pulse in the time domain, which is a linear superposition of plane waves with different frequencies, will be distorted if the dispersion introduced by the finite-difference scheme is intolerable. Therefore, for simplicity, we will analyze the grid-dispersion error in a homogeneous free space medium.

To ascertain the grid-dispersion error, we assume a time-harmonic solution where the Fourier mode has time dependence of the form $A_0 e^{-i\omega l\Delta t}$ in (37.1.20). In this case, the left-hand side of (37.1.24) becomes

$$\left(e^{-i\omega\Delta t} - 2 + e^{+i\omega\Delta t}\right)\phi^l_{m,n,p} = -4\sin^2\left(\frac{\omega\Delta t}{2}\right)\phi^l_{m,n,p}. \tag{37.1.30}$$

Then, from (37.1.24), it follows that

$$\sin\left(\frac{\omega\Delta t}{2}\right) = rs, \tag{37.1.31}$$

where r and $s(k_x, k_y, k_z)$ are given in (37.1.25). Now, (37.1.31) governs the relationship between ω and k_x, k_y, and k_z in the finite-difference scheme, and hence, is a dispersion relation for the finite-difference approximate solution.

The above gives a rather complicated relationship between the frequency ω and the wave numbers k_x, k_y, and k_z. This is the result of the finite-difference approximation of the scalar wave equation. As a sanity check, when the space and time discretizations become very small, we should recover the dispersion relation of homogeneous medium or free space.

But if a medium is homogeneous, it is well known that (37.1.1) has a plane-wave solution of the type given by (37.1.19) where

$$\omega = c\sqrt{k_x^2 + k_y^2 + k_z^2} = c|\mathbf{k}| = ck. \tag{37.1.32}$$

where $\mathbf{k} = \hat{x}k_x + \hat{y}k_y + \hat{z}k_z$ is the direction of propagation of the plane wave. Defining the phase velocity to be $\omega/k = c$, this phase velocity is isotropic, or the same in all directions. Moreover, it is independent of frequency.

But in (37.1.31), because of the definition of s as given by (37.1.25), the dispersion relation between ω and $\mathbf{k}$ is not isotropic (anisotropic). This implies that plane waves propagating in different directions will have different phase velocities.

Equation (37.1.31) is the dispersion relation for the approximate solution. It departs from Equation (37.1.32), the exact dispersion relation for free space, due to the finite-difference approximation. This departure gives rise to errors called grid dispersion errors. For example, when c is a constant, (37.1.32) states that the phase velocities of plane waves of different wavelengths and directions are the same. However, this is not true for (37.1.31), as shall be shown.

To elaborate more on the grid dispersion error, we assume that s is small. Then (37.1.31), after using Taylor series expansion, can be written approximately as

$$\frac{\omega\Delta t}{2} = \sin^{-1} rs \cong rs + \frac{r^3 s^3}{6}. \tag{37.1.33}$$

When Δs is small, using the small argument approximation for the sine function, one obtains from (37.1.25)

$$s \simeq \frac{\Delta s}{2}(k_x^2 + k_y^2 + k_z^2)^{1/2}\left[1 - \frac{\Delta s^2}{24}\left(\frac{k_x^4 + k_y^4 + k_z^4}{k_x^2 + k_y^2 + k_z^2}\right)\right] \tag{37.1.34}$$

Equation (37.1.33), by taking its higher-order Taylor expansion, then becomes

$$\frac{\omega\Delta t}{2} \simeq r\frac{\Delta s}{2}(k_x^2 + k_y^2 + k_z^2)^{1/2}\,[1 - \delta] \tag{37.1.35}$$

where (see [37])

$$\delta = \frac{\Delta s^2}{24}\frac{k_x^4 + k_y^4 + k_z^4}{k_x^2 + k_y^2 + k_z^2} - \frac{r^2\Delta s^2}{24}(k_x^2 + k_y^2 + k_z^2) \tag{37.1.36}$$

It has to be reminded that $r = c\Delta t/\Delta s$, a dimensionless quantity. From the above, (37.1.35) is almost the same as (37.1.32) save for the factor $1 - \delta$. Also, if $\delta = 0$, we retrieve the dispersion relation of the homogeneous free-space medium. So δ is a measure of the departure of the dispersion relation from that of free space due to our finite-difference approximation. More insight can be gotten if we let $(k_x, k_y, k_z) = (k\sin\theta\cos\phi, k\sin\theta\sin\phi, k\cos\theta)$, then the above can be rewritten as

$$\delta = \frac{k^2\Delta s^2}{24}F(\theta,\phi) - \frac{r^2\Delta k^2\Delta s^2}{24}G(\theta,\phi) \tag{37.1.37}$$

where $F(\theta,\phi)$ and $G(\theta,\phi)$ are functions of $O(1)$ that depends only on θ,ϕ. Since r^2 is less than one, the dominant term above is the first term. If $\frac{k^2\Delta s^2}{24} \ll 1$, then δ small, reducing the grid-dispersion error.

Since $\mathbf{k}$ is inversely proportional to wavelength λ, then δ in the correction to the above equation is proportional to $2\pi\Delta s^2/\lambda^2$. Therefore, to reduce the grid dispersion error, it is necessary for δ to be small or to have

$$\frac{1}{24}\left(\frac{2\pi\Delta s}{\lambda}\right)^2 \ll 1. \tag{37.1.38}$$

Or the space discretization Δs has to be smaller than the wavelength in question to mitigate the grid-dispersion error. When this is true, using the fact that $r = c\Delta t/\Delta s$, then (37.1.35) becomes

$$\frac{\omega}{c} \approx \sqrt{k_x^2 + k_y^2 + k_z^2}\,. \tag{37.1.39}$$

which is close to the dispersion relation of free space as indicated in (37.1.32). Furthermore, Δt must be chosen so that the CFL stability criterion is met. Therefore, the rule of thumb is to choose about 10 to 20 grid points per wavelength. Also, for a plane wave propagating as $e^{i\mathbf{k}\cdot\mathbf{r}}$, an error $\delta\mathbf{k}$ in the vector $\mathbf{k}$ gives rise to cumulative error $e^{i\delta\mathbf{k}\cdot\mathbf{r}}$. The larger the distance traveled, the larger is $\mathbf{r}$, and the larger the cumulative phase error. And hence, the grid size must be smaller in order to arrest such phase error due to the grid dispersion!

37.2 The Yee Algorithm

The Yee algorithm (Yee 1966 [287])[4] is a simple algorithm specially designed to solve vector electromagnetic field problems on a rectilinear grid. The finite-difference time-domain (FDTD) method (Taflov 1988) when applied to solving electromagnetics problems, usually uses this algorithm. To derive it, Maxwell's equations in the time-domain are first written in cartesian coordinates:

$$-\frac{\partial B_x}{\partial t} = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}, \tag{37.2.1}$$

$$-\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}, \tag{37.2.2}$$

$$-\frac{\partial B_z}{\partial t} = \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}, \tag{37.2.3}$$

$$\frac{\partial D_x}{\partial t} = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} - J_x, \tag{37.2.4}$$

$$\frac{\partial D_y}{\partial t} = \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} - J_y, \tag{37.2.5}$$

$$\frac{\partial D_z}{\partial t} = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} - J_z. \tag{37.2.6}$$

[4]Note that this algorithm, together with the method of moments [191] for solving Maxwell's equations, emerge shortly after the advent of IC based digital computers.

Before proceeding any further, it is prudent to rewrite the differential equation form of Maxwell's equations in their integral form. For instance, the first equation above can be rewritten as[5]

$$-\frac{\partial}{\partial t}\iint_{\Delta S} B_x dS = \oint_{\Delta C} \mathbf{E}\cdot d\mathbf{l} \tag{37.2.7}$$

where $\Delta S = \Delta x \Delta z$. The approximation of this integral form will be applied to the face that is closest to the observer in Figure 37.3. Hence, one can see that the curl of $\mathbf{E}$ is proportional to the time-derivative of the magnetic flux through the suface enclosed by ΔC, which is ΔS.

One can see this relationship for the other surfaces of the cube in the figure as well: the electric field is curling around the magnetic flux. For the second half of the above equations, one can see that the magnetic fields are curling around the electric flux, but on a staggered grid. These two staggered grids are intertwined with respect to each other. This is the spirit with which the Yee algorithm is written. He was apparently motivated by fluid dynamics when he did the work.

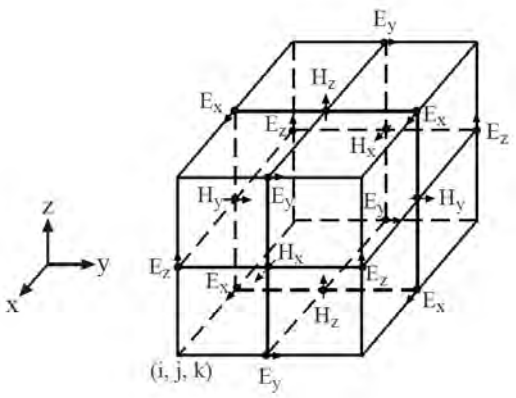


Figure 37.3: The assignment of fields on a grid in the Yee algorithm [287]. This algorithm is vastly popular for electromagnetic simulations [289].

After denoting $f(m\Delta x, n\Delta y, p\Delta z, l\Delta t) = f^l_{m,n,p}$, a more compact notation, where the superscript is time, and subscripts are space points, and replacing derivatives with central finite-differences in accordance with Figure 37.3, (37.2.1) becomes

$$\begin{aligned}\frac{1}{\Delta t}\left[B^{l+\frac{1}{2}}_{x,m,n+\frac{1}{2},p+\frac{1}{2}} - B^{l-\frac{1}{2}}_{x,m,n+\frac{1}{2},p+\frac{1}{2}}\right] = \frac{1}{\Delta z}\left[E^l_{y,m,n+\frac{1}{2},p+1} - E^l_{y,m,n+\frac{1}{2},p}\right] \\ -\frac{1}{\Delta y}\left[E^l_{z,m,n+1,p+\frac{1}{2}} - E^l_{z,m,n,p+\frac{1}{2}}\right].\end{aligned} \tag{37.2.8}$$

where the above formula is evaluated at $t = l\Delta t$. Moreover, the above can be repeated for (37.2.2) and (37.2.3). Notice that in Figure 37.3, the electric field is always assigned to the edge center of a cube, whereas the magnetic field is always assigned to the face center of a cube.[6]

[5]The finite integration technique developed by T. Weiland is also of note [298, 299].

[6]This algorithm is intimately related to differential forms which has given rise to the area of discrete exterior calculus [300].

In fact, after multiplying (37.2.8) by $\Delta z \Delta y$, (37.2.8) is also the approximation of the integral forms of Maxwell's equations when applied at a face of a cube. By doing so, the left-hand side of (37.2.8), by (37.2.7), becomes

$$(\Delta y \Delta z/\Delta t)\left[B^{l+\frac{1}{2}}_{x,m,n+\frac{1}{2},p+\frac{1}{2}} - B^{l-\frac{1}{2}}_{x,m,n+\frac{1}{2},p+\frac{1}{2}}\right], \tag{37.2.9}$$

which is the time variation of the total flux through an elemental area $\Delta y \Delta z$. Moreover, by summing this flux on the six faces of the cube shown in Figure 37.3, and using the right-hand side of (37.2.8) and its equivalence, it can be shown that the magnetic flux adds up to zero. Hence, $\frac{\partial}{\partial t}\nabla\cdot\mathbf{B} = 0$ condition is satisfied within the numerical approximations of Yee algorithm. The above shows that if the initial value implies that $\nabla \cdot \mathbf{B} = 0$, the algorithm will preserve this condition. So even though we are solving Faraday's law, Gauss' law is also enforced if the cumulative numerical error is kept small. This is important in maintaining the stability of the numerical algorithm [37].

Furthermore, a similar approximation of (37.2.4) leads to

$$\begin{aligned}\frac{1}{\Delta t}\left[D^{l}_{x,m+\frac{1}{2},n,p} - D^{l-1}_{x,m+\frac{1}{2},n,p}\right] &= \frac{1}{\Delta y}\left[H^{l-\frac{1}{2}}_{z,m+\frac{1}{2},n+\frac{1}{2},p} - H^{l-\frac{1}{2}}_{z,m+\frac{1}{2},n-\frac{1}{2},p}\right] \\ &\quad -\frac{1}{\Delta z}\left[H^{l-\frac{1}{2}}_{y,m+\frac{1}{2},n,p+\frac{1}{2}} - H^{l-\frac{1}{2}}_{y,m+\frac{1}{2},n,p-\frac{1}{2}}\right] - J^{l-\frac{1}{2}}_{x,m+\frac{1}{2},n,p}.\end{aligned} \tag{37.2.10}$$

By the same token, similar approximations apply for (37.2.5) and (37.2.6). In addition, the above has an interpretation similar to (37.2.8) if one thinks in terms of a cube that is shifted by half a grid point in each direction to form a staggered grid. Hence, the approximations of (37.2.4) to (37.2.6) are consistent with the approximation of $\frac{\partial}{\partial t}\nabla \cdot \mathbf{D} = -\nabla \cdot \mathbf{J}$. This manner of alternatively solving for the $\mathbf{B}$ and $\mathbf{D}$ fields in tandem while the fields are placed on a staggered grid is also called the leap-frog scheme.

In the above, $\mathbf{D} = \epsilon\mathbf{E}$ and $\mathbf{B} = \mu\mathbf{H}$. Since the magnetic field and the electric field are assigned on staggered grids, μ and ϵ may have to be assigned on staggered grids. This does not usually lead to serious problems if the grid size is small. Alternatively, (37.1.16) and (37.1.17) can be used to remove this problem, and to reduce storage.

By eliminating the $\mathbf{E}$ or the $\mathbf{H}$ field from the Yee algorithm, it can be shown that the Yee algorithm is equivalent to finite differencing the vector wave equation directly. Hence, the Yee algorithm is also constrained by the CFL stability criterion [295].

The following figures show some results of FDTD simulations. Because the answers are in the time-domain, beautiful animations of the fields are also available online:

https://www.remcom.com/xfdtd-3d-em-simulation-software

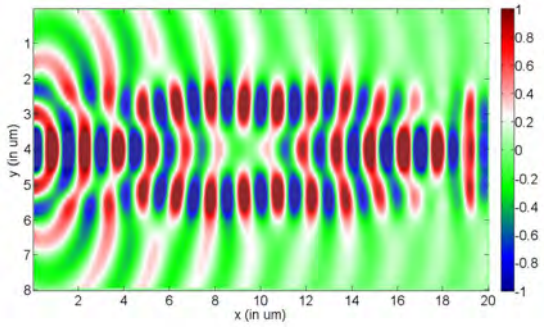


Figure 37.4: The 2D FDTD simulation of complicated optical waveguides. Such simulations can be done from static to optical frequencies (courtesy of Mathworks).

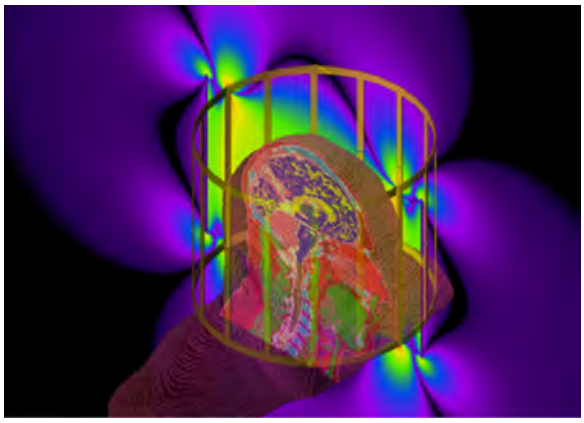


Figure 37.5: FDTD simulation of human head in a squirrel cage of an MRI (magnetic resonance imaging) system. A static magnetic field biases the spins in the human body. Then an RF field is used to tilt the spins causing them to precess. Their precession gives rise to electromagnetic radiation (also called spin echo) that can be measured by the squirrel cage coils (courtesy of REMCOM).

37.2.1 Finite-Difference Frequency Domain Method

Unlike electrical engineering, in many fields, nonlinear problems are prevalent. But when we have a linear time-invariant problem, it is simpler to solve the problem in the frequency domain. This is analogous to perform a time Fourier transform of the pertinent linear equations.

Consequently, one can write (37.2.1) to (37.2.6) in the frequency domain to remove the time derivatives. Then one can apply the finite difference approximation to the space derivatives using the Yee grid. As a result, in replacement of Maxwell's equations, one arrives at a matrix equation

$$\overline{\mathbf{A}} \cdot \mathbf{x} = \mathbf{b} \tag{37.2.11}$$

where $\mathbf{x}$ is an unknown vector containing $\mathbf{E}$ and $\mathbf{H}$ fields, and $\mathbf{b}$ is a source vector that drives the system containing $\mathbf{J}$. The above matrix-vector product can be effected using the Yee algorithm. Due to the near-neighbor interactions of the fields on the Yee grid, the matrix $\overline{\mathbf{A}}$ is highly sparse and contains $O(N)$ non-zero elements. When an iterative method is used to solve the above equation, the major cost is in performing a matrix-vector product $\overline{\mathbf{A}} \cdot \mathbf{x}$. However, in practice, the matrix $\overline{\mathbf{A}}$ is never generated nor stored making this a matrix-free method. Because of the simplicity of the Yee algorithm, a code can be easily written to produce the action of $\overline{\mathbf{A}}$ on $\mathbf{x}$ or $\overline{\mathbf{A}} \cdot \mathbf{x}$.

37.3 Absorbing Boundary Conditions

It will not be complete to close this lecture without mentioning absorbing boundary conditions. As computer has finite memory, space of infinitely large extent cannot be simulated with finite computer memory. Hence, it is important to design absorbing boundary conditions at the walls of the simulation domain or box, so that waves impinging on them are not reflected. This mimicks the physics of an infinitely large box.

This is analogous to experiments in microwave engineering. In order to perform experiments in an infinite space, such experiments are usually done in an anechoic (non-echoing or non-reflecting) chamber. An anechoic chamber has its walls padded with absorbing materials or microwave absorbers so as to minimize the reflections off its walls (see Figure 37.6). The electromagnetically quiet environment is important for studying EMC/EMI (electromagnetics compatibility/electromagnetics interference) problems. This is an important problem in many industries that use electromagnetics technologies. Figure 37.7 shows an acoustic equivalence of anechoic chamber.

Figure 37.6: An anechoic chamber for radio frequency. In such an electromagnetically quiet chamber, interference from other RF equipment is minimized (courtesy of Panasonic).

Figure 37.7: An acoustic anechoic chamber. In such a chamber, there is no reflection from the walls of the chamber; even the breast-feeding sound of a baby can be heard clearly (courtesy of AGH University, Poland).

By the same token, in order to simulate numerically an infinitely large box with a finite-size box, absorbing boundary conditions (ABCs) are designed at its walls. The simplest of such ABCs is the impedance boundary condition. (A transmission line terminated with an impedance reflects less than one terminated with an open or a short circuit.) Another simple ABC is to mimick the Sommerfeld radiation condition (much of this is reviewed in [37]).[7]

A recently invented ABC is the perfectly matched layers (PML) [301]. Also, another similar ABC is the stretched coordinates PML [302]. Figure 37.8 shows simulation results with and without stretched coordinates PMLs on the walls of the simulation domain [303].

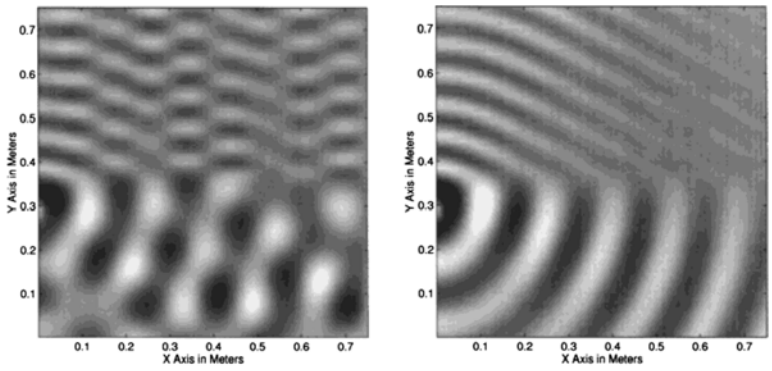


Figure 37.8: Simulation of a source on top of a half-space (left) without stretched coordinates PML ABC; and (right) with stretched coordinats PML ABC [303]. One can see the effect standing wave patterns due to the reflections from the walls of the simulation domain. The reflected waves give rise to interference patterns that are clearly visible in the left figure. They disappear upon the introduction of stretched coordinates PML ABC.

[7] ABCs are beyond the scope of these lecture notes.

Exercises for Lecture 37

Problem 37-1: This problem refers to Chapter 37.

(i) Show that (37.1.14) expands to (37.1.12) after using (37.1.13) and its equivalence in x and y directions. Hence, derive (37.1.18).

(ii) Show that (37.1.28) is indeed true. Give the physical meaning of (37.1.29).

(iii) Derive (37.1.35) and (37.1.36). What is the physical meaning of the extra δ in (37.1.35)?

Chapter 38

Quantum Theory of Electromagnetics

The quantum theory of the world is the culmination of a series of astonishing intellectual exercises. It is often termed the intellectual triumph of the twentieth century. One often says that deciphering the laws of nature is like watching two persons play a chess game with rules unbeknownst to us. By watching the moves, conjectures were made about the rules, which were later confirmed by experiments. After much sweat and tears, we finally have pieced together these perplexing rules. But we are grateful that these laws of nature are revealed to us, and on top of them we can build new technologies with them.

It is important to know that the quantum theory of electromagnetics[1] emerges alongside with quantum theory. This new quantum theory of electromagnetics is intimately related to Maxwell's equations as shall be seen. It also inspired quantum field theory [45, 304]. This new theory spawns the possibility for quantum technologies, one of which is quantum computing. Others are quantum communication, quantum cryptography, quantum sensing, and many more.

A recent major advancement in quantum theory is the confirmation of the correctness of quantum interpretation. This was not done until 1982 and it ushers in the new era of quantum information science. In 2022, the Nobel Prize in physics was given to Alain Aspect [305], John F. Clauser [306], and Anton Zeilinger [307] for their contributions to quantum information science.

38.1 Historical Background on Quantum Theory

Initially, light was thought to be made of particles. That light is a wave was demonstrated by Newton's ring phenomenon [21] in the eighteenth century (1717) (see Figure 38.1). In 1801, Thomas Young demonstrated the double slit experiment for light [308] that further confirmed its wave nature (see Figure 38.2). But by the beginning of the 20-th century, one has to accept

[1] This is often called quantum theory of light in many textbooks. This is because this theory is first manifested in optical frequencies with optical photons. Nowadays, microwave photons are detectable in laboratory, and it will come a day that this quantum theory will manifest itself broadly through the electromagnetic spectrum.

that light is both a particle, called a photon, carrying a quantum of energy with a quantum of momentum, as well as a particle endowed with wave-like behavior. This is called wave-particle duality. We shall outline the historical reason for this development.

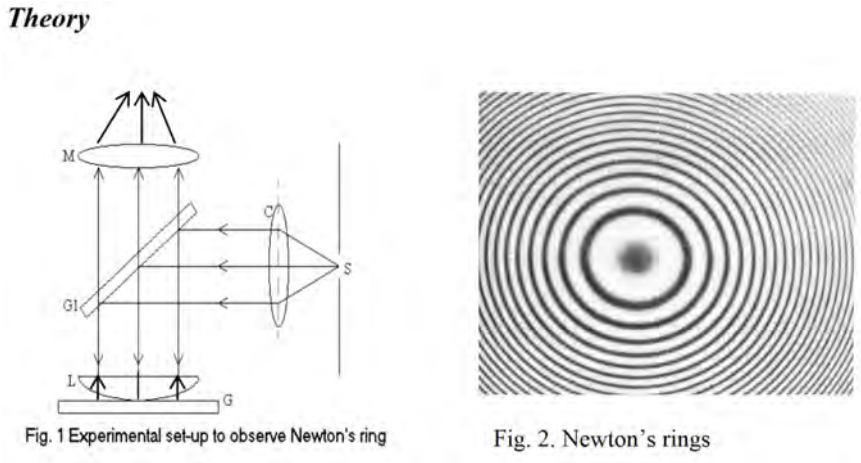


Figure 38.1: A Newton's rings experiment that indicates the wave nature of light (courtesy of [309]).

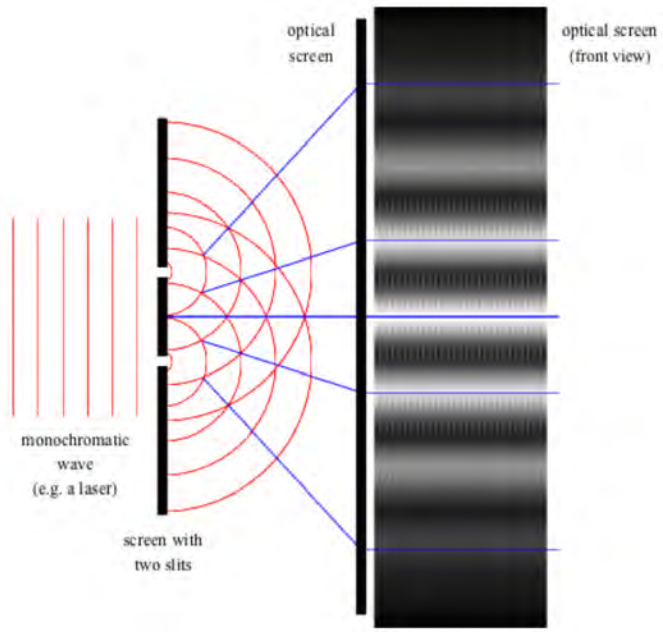


Figure 38.2: A Young's double-slit experiment. Again, the interference pattern reveals the wave nature of light (courtesy of [310]).

As mentioned above, quantum theory is a major intellectual achievement of the twentieth century, even though new knowledge is still emerging in it. Several major experimental findings led to the revelation of quantum theory of nature.

In nature, we know that matter is not infinitely divisible. This is vindicated by the atomic theory of John Dalton (1766-1844) [311]. So fluid is not infinitely divisible: as when water is divided into smaller pieces, one will eventually arrive at water molecule, H_2O, which is the fundamental building block of water.

It turns out that electromagnetic energy is not infinitely divisible either. The electromagnetic radiation out of a heated cavity would have a very different spectrum if electromagnetic energy is infinitely divisible. In order to fit experimental observation of radiation from a heated electromagnetic cavity, Max Planck (1900s) [312] proposed that electromagnetic energy comes in packets or is quantized. Each packet of energy or a quantum of energy E is associated with the frequency of electromagnetic wave, namely

$$E = \hbar\omega = \hbar 2\pi f = hf \tag{38.1.1}$$

where $\hbar = h/(2\pi)$ is now known as the reduced Planck constant and $h = 6.626 \times 10^{-34}$ J·s (Joule-second). Since $\hbar$ is very small, this packet of energy is very small unless ω is large. So it is no surprise that the quantization of electromagnetic field is first associated with light, a very high frequency electromagnetic radiation. A red-light photon at a wavelength of 700 nm corresponds to an energy of approximately 2 eV $\approx 3 \times 10^{-19}$J $\approx 75\, k_B T$, (where $k_B T$ denotes the thermal energy

from thermal law, and k_B is Boltzmann's constant. This is about 25 meV at room temperature.[2]) A microwave photon has approximately 1×10^{-5} eV $\approx 10^{-2}$meV, making it a lot harder to detect compared to an optical photon.

The second experimental evidence that light is quantized is the photo-electric effect [313]. It was found that matter emitted electrons when light shined on it. First, the light frequency has to correspond to the "resonant" frequency of the atom.[3] Second, the number of electrons emitted is proportional to the number of packets of energy $\hbar\omega$ that the light carries. This was a clear indication that light energy traveled in packets or quanta as posited by Einstein in 1905.

This wave-particle duality concept mentioned at the beginning of this section was not new to quantum theory as electrons were known to behave both like a particle and a wave. The particle nature of an electron was confirmed by the measurement of its charge by Millikan in 1913 in his oil-drop experiment. (The double slit experiment for electron was done in 1927 by Davison and Germer, indicating that an electron has a wave nature as well [308].) In 1924, De Broglie [88] suggested that there is a wave associated with an electron with momentum p such that

$$p = \hbar k \tag{38.1.2}$$

where $k = 2\pi/\lambda$, the wavenumber. All this knowledge gave hint to the quantum theorists of that era to come up with a new way to describe nature.

Classically, particles like an electron moves through space obeying Newton's laws of motion first established in 1687 [314]. The old way of describing particle motion is known as classical mechanics, but the new way of describing particle motion is known as quantum theory. Quantum theory is very much motivated by a branch of classical mechanics called Hamiltonian theory. We will first use Hamiltonian theory to study a simple pendulum and connect it with electromagnetic oscillations.

[2]This is a number ought to be remembered by semi-conductor scientists as the size of the material bandgap with respect to this thermal energy decides if a material is a semi-conductor, conductor, or insulator at room temperature.

[3]This is akin to the physics of resonant tunneling in antennas (see Figure 31.4).

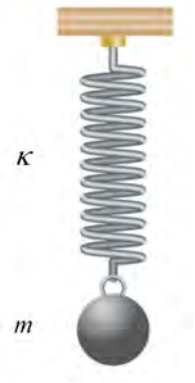


Figure 38.3: A classical pendulum can be used to illustrate classical Hamiltonian theory. The equation of motion of a classical pendulum can be derived using Newton's law or Hamiltonian theory. As such, it is easier to illustrate quantum theory of a pendulum starting with classical Hamiltonian theory. In the figure, κ and m are spring constant and mass, respectively of the pendulum (courtesy of pngwing.com).

Consider a time-harmonic electromagnetic oscillation: it can be the oscillation of a cavity mode, or that of a travelling electromagnetic wave. If we were to gaze at the electromagnetic oscillation and observe the field at the location z, we observe a simple harmonic oscillation resembling that of a pendulum. Hence, we will study a simple pendulum first, and discuss the quantum theory behind the simple pendulum. Hence, we will then connect electromagnetic oscillations to the quantum pendulum. (We can think of the macroscopic electromagnetic oscillations observed in a cavity, even in vacuum, is the cooperative effect of electromagnetic oscillations at the root level caused possibly by electron-positron oscillations.)

A Simple Pendulum

As we have seen in the Drude-Lorentz-Sommerfeld mode, for a particle of mass m attached to a spring connected to a wall, where the restoring force is like Hooke's law, the equation of motion of a pendulum by Newton's law is

$$m\frac{d^2q}{dt^2} + \kappa q = 0 \tag{38.1.3}$$

where κ is the spring constant, and we assume that the oscillator is not driven by an external force, but is in natural or free oscillation. Here, q is used to denote the displacement of the pendulum from its quiescent position. The above equation is homomorphic/analgous to a cavity resonance problem. We can see that $q \Leftrightarrow E_0$ relates the two equations. By letting[4]

$$q = q_0 e^{-i\omega t} \tag{38.1.4}$$

[4]As aforementioned, for this part of the lecture, we will switch to using $\exp(-i\omega t)$ time convention as is commonly used in optics and physics literatures.

where q_0 is the amplitude of this complex signal, the above (38.1.3) becomes[5]

$$-m\omega^2 q_0 + \kappa q_0 = 0 \tag{38.1.5}$$

Again, a non-trivial solution is possible only at the resonant frequency of the oscillator or that when $\omega = \omega_0$ where

$$\omega_0 = \sqrt{\frac{\kappa}{m}} \tag{38.1.6}$$

This is the natural solution or resonant solution of the system.[6]

38.2 Hamiltonian Theory

Quantum theory is initimately related to Hamiltonian theory. Equation (38.1.3) can be derived by Newton's law but it is more interesting, as an alternative, to derive it via Hamiltonian theory. This is because Hamiltonian theory motivates quantum theory.

Hamiltonian theory, developed by Hamilton (1805-1865) [315], is motivated by energy conservation [316]. The Hamiltonian H of a system is given by its total energy, namely that

$$H = T + V \tag{38.2.1}$$

where T is the kinetic energy and V is the potential energy of the system.

For a simple pendulum, the kinetic energy is given by

$$T = \frac{mv^2}{2} = \frac{m^2v^2}{2m} = \frac{p^2}{2m} \tag{38.2.2}$$

where $p = mv$ is the momentum of the particle. The potential energy, assuming that the particle is attached to a spring with spring constant κ and displaced q from equilibrium, is given by

$$V = \frac{1}{2}\kappa q^2 = \frac{1}{2}m\omega_0^2 q^2 \tag{38.2.3}$$

Hence, the Hamiltonian is given by

$$H = T + V = \frac{p^2}{2m} + \frac{1}{2}m\omega_0^2 q^2 \tag{38.2.4}$$

[5]In optics, instead of using phasors as in electrical engineering, it is a custom to use a complex signal with the understanding that the real part of this complex signal is time-harmonic similar to the phasor technique [60], Whenever a real signal is needed, one adds the 'complex conjugate' (c.c.) part. Hence, q_0 is not necessarily real-valued. Thus, this does not imply the q and E_0 are equal to each other. In this case, q_0 is a complex amplitude, while E_0 is real-valued amplitude because of definition of the phasor technique. These two values, q_0 and E_0 are analogous to each other and used by two different communities to denote the amplitude of a signal. What cultural diversities we have even within finding the solutions to Maxwell's equations!

[6]As aforementioned, natural solution, resonant solution, homogeneous solution (use in ODE), and null-space solution (linear algebra) are similar concepts.

At any instant of time t, we assume that $p(t) = mv(t) = m\frac{d}{dt}q(t)$ is independent of $q(t)$.[7] In other words, they can vary independently of each other. But $p(t)$ and $q(t)$ have to time evolve to conserve energy or to keep H, the total energy, constant or independent of time. In other words,

$$\frac{d}{dt}H\left(p(t), q(t)\right) = 0 = \frac{dp}{dt}\frac{\partial H}{\partial p} + \frac{dq}{dt}\frac{\partial H}{\partial q} \tag{38.2.5}$$

Therefore, the Hamilton equations of motion are derived to be[8]

$$\frac{dp}{dt} = -\frac{\partial H}{\partial q}, \qquad \frac{dq}{dt} = \frac{\partial H}{\partial p} \tag{38.2.6}$$

From (38.2.4), we gather that

$$\frac{\partial H}{\partial q} = m\omega_0^2 q, \qquad \frac{\partial H}{\partial p} = \frac{p}{m} \tag{38.2.7}$$

Applying (38.2.6), we have[9]

$$\frac{dp}{dt} = -m\omega_0^2 q, \qquad \frac{dq}{dt} = \frac{p}{m} \tag{38.2.8}$$

Combining the two equations in (38.2.8) above, we have

$$m\frac{d^2q}{dt^2} = -m\omega_0^2 q = -\kappa q \tag{38.2.9}$$

which is also derivable by Newton's law.

A typical harmonic oscillator solution to (38.2.9) is

$$q(t) = q_0 \cos(\omega_0 t + \psi) \tag{38.2.10}$$

The corresponding momentum $p(t) = m\frac{dq}{dt}$ is

$$p(t) = -mq_0\omega_0 \sin(\omega_0 t + \psi) \tag{38.2.11}$$

Hence, one can easily show that the Hamiltonian H is a constant by

$$\begin{aligned} H &= \frac{1}{2}m\omega_0^2 q_0^2 \sin^2(\omega_0 t + \psi) + \frac{1}{2}m\omega_0^2 q_0^2 \cos^2(\omega_0 t + \psi) \\ &= \frac{1}{2}m\omega_0^2 q_0^2 = E \end{aligned} \tag{38.2.12}$$

And the total energy E is a constant of motion (physicists parlance for a time-independent variable); it depends only on the amplitude q_0 of the oscillation in (38.2.10).

[7] $p(t)$ and $q(t)$ are termed conjugate variables in many textbooks.

[8] Note that the Hamilton equations are determined to within a multiplicative constant, because one has not stipulated the connection between space and time, or we have not calibrated our clock [316].

[9] We can also calibrate our clock here so that it agrees with our definition of momentum in the ensuing equation.

38.3 Schrödinger Equation (1925)

Having seen the Hamiltonian theory for describing a simple pendulum which is homomorphic to an electromagnetic oscillation, we shall next see the quantum theory description of the same simple pendulum, using it to inspire quantum electromagnetics oscillations.

Schrödinger equation cannot be derived: It is a wonderful result of a postulate and a powerful guessing game based on experimental observations [77, 78].[10] Hamiltonian theory says that

$$H = \frac{p^2}{2m} + \frac{1}{2}m\omega_0^2 q^2 = E \tag{38.3.1}$$

where E is the total energy of the oscillator, or pendulum.

To build this randomness into a quantum harmonic oscillator, a new theory is needed. The position of the particle is described by the wave function,[11] which makes the location of the particle uncertain. To this end, Schrödinger proposed his equation which is a partial differential equation. He was very much motivated by the experimental revelation then that $p = \hbar k$ from De Broglie [88] and that $E = \hbar\omega$ from Planck's law [24] and the photo-electric effect. Equation (38.3.1) can be rewritten more suggestively as

$$\frac{\hbar^2 k^2}{2m} + \frac{1}{2}m\omega_0^2 q^2 = \hbar\omega \tag{38.3.2}$$

We can associate $ik = \frac{\partial}{\partial q}$ and $-i\omega = \frac{\partial}{\partial t}$ as they will become such when operating on the time-harmonic plane wave $e^{ikq-i\omega t}$. To add more texture to the above equation, one lets the above become an operator equation that operates on a wave function $\Psi(q,t)$ so that

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial q^2}\Psi(q,t) + \frac{1}{2}m\omega_0^2 q^2\Psi(q,t) = i\hbar\frac{\partial}{\partial t}\Psi(q,t) \tag{38.3.3}$$

The wave function $\Psi(q,t)$ above is used to describe the state of the particle giving it a wave nature.[12] As aforementioned, if the wave function is of the form

$$\Psi(q,t) \sim e^{ikq-i\omega t} \tag{38.3.4}$$

then upon substituting (38.3.4) back into (38.3.3), we retrieve (38.3.2).

Equation (38.3.3) is Schrödinger equation (or the Schrödinger wave equation) in one dimension for the quantum version of the simple harmonic oscillator. In Schrödinger equation, we can further posit that the wave function has the general form

$$\Psi(q,t) = e^{ikq-i\omega t}A(q,t) \tag{38.3.5}$$

[10] Rumour has it that he got the inspiration after he went into the mountain for a retreat.

[11] Since a function is equivalent to a vector (see Section 36.4), and this wave function describes the state of the quantum system such as a quantum pendulum, this is also called a state vector.

[12] The term "state" may have been inherited from the control theorist. Thinking of a function as a vector, Schrödinger equation reminds us of the state variable approach in control theory where the state of a system is described by a state variable vector $\mathbf{v}(t)$. The time evolution of the state variable in control theory, the simplest version, is $d\mathbf{v}(t)/dt = \overline{\mathbf{A}} \cdot \mathbf{v}(t)$. This is homomorphic to Schrödinger equation.

where $A(q,t)$ is a slowly varying function of q and t, compared to $e^{ikq-i\omega t}$.[13] In other words, this is the expression for a wave packet. With this wave packet, the $\partial^2/\partial q^2$ can be again approximated by $-k^2$ in the short-wavelength limit, or when $k \to \infty$, as has been done in the paraxial wave approximation (see Sect. 33.3.1). Furthermore, if the wave function is assumed to be quasi-monochromatic, then $i\hbar\partial/\partial_t \Psi(q,t) \approx \hbar\omega\Psi(q,t)$, we again retrieve the classical equation in (38.3.2) from (38.3.3). Hence, the classical equation (38.3.2) is a short wavelength, monochromatic approximation of Schrödinger equation for a wave packet. (However, as we shall see, the solutions to Schrödinger equation are not limited to just wave packets described by (38.3.5).)

Correspondence Principle

In the limit when $\hbar \to 0$, the quantization energy will be very small, and we expect to retrieve the classical picture or classical mechanics. In fact, when $\hbar \to 0$, if the particle is to have a finite amount of energy E, the frequency $\omega \to \infty$. Consequently, for a particle carrying a finite momentum, $k \to \infty$ as well. Therefore, the wave function $\Psi(q,t)$ becomes a very high-frequency wave function or a wave packet. One can see that this wave packet follows the classical equations of motion. This is gist of the correspondence principle [317].

Wave functions

In classical mechanics, the position of a particle is described by the variable q, but in the quantum world, the position of a particle q is elevated to be an operator associated with a random variable. There has to be a probability density function associated with this random variable. We shall learn that this property is related to the wave function that is the solution to Schrödinger equation, which we will study in detail next. As aforementioned, this is also known as wave-particle duality.

Stationary Solution and Eigenvalue Problem

The above can be converted into an eigenvalue problem, just as in waveguide and cavity problems, using separation of variables, by letting[14]

$$\Psi(q,t) = \Psi_n(q)e^{-i\omega_n t} \tag{38.3.6}$$

By so doing, (38.3.3) becomes an eigenvalue problem

$$\left[-\frac{\hbar^2}{2m}\frac{d^2}{dq^2} + \frac{1}{2}m\omega_0^2 q^2\right]\Psi_n(q) = E_n\Psi_n(q) \tag{38.3.7}$$

where $E_n = \hbar\omega_n$ is the eigenvalue while $\Psi_n(q)$ is the corresponding eigenfunction.

The parabolic q^2 potential profile is also known as a quadratic potential well as it provides the restoring force to keep the particle bound to the well classically (see Section 38.2 and (38.2.8)). (The above equation is also similar to the electromagnetic equation for a dielectric slab waveguide, where the second term is a dielectric profile that can trap a waveguide mode. Therefore, the potential well is a trap for the particle both in classical mechanics as expressed in (38.2.4) or in

[13] Recall that this is similar in spirit when we study high frequency solutions of Maxwell's equations and paraxial wave approximation in Sect. 33.3.1.

[14] Mind you, the following is ω_n, not ω_0.

wave physics as in (38.3.7). It keeps the particle bound to the well classically (see Section 38.2 and (38.2.8)).

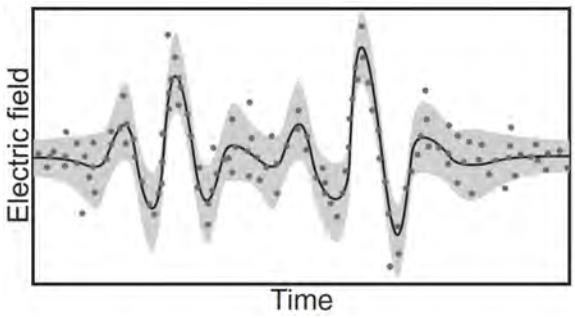


Figure 38.4: Schematic representation of the randomness of the measured electric field. The electric field amplitude is a quantum observable, and hence, is a random variable. The electric field amplitude maps to the displacement (position) of the quantum pendulum, which is also a random variable. The dots represent the one possible outcome of the measured value at the given time of the observable in the laboratory. Because the observable is a random variable, the measured value is not the same at a given time if the measurement is repeated. The shaded zone indicates the range of possible values of the observable since this is a random variable. The solid line indicates the average (expectation) value of the observable after averaging over many measurements. (courtesy of Kira and Koch [318]).

The above equation (38.3.7) can be solved in closed form in terms of Hermite-Gaussian functions (1864) [319], or that[15]

$$\Psi_n(q) = \sqrt{\frac{1}{2^n n!}\sqrt{\frac{m\omega_0}{\pi\hbar}}} e^{-\frac{m\omega_0}{2\hbar}q^2} H_n\left(\sqrt{\frac{m\omega_0}{\hbar}}q\right) \tag{38.3.8}$$

where $H_n(y)$[16] is a Hermite polynomial, and the eigenvalues are found in closed form as well, viz.,[17]

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega_0 \tag{38.3.9}$$

Here, the eigenfunction or eigenstate $\Psi_n(q)$ is known as the photon number state (or just a number state or Fock state in short) of the solution. It corresponds to having n "photons" in the oscillation.

[15]Lucky that we are, all these were figured out by contemporaries of James Clerk Maxwell. We stand on the shoulders of giants!

[16]The mass m is that of the particle that forms the quantum pendulum. It could be an electron, an atom, or some other particles.

[17]It is to be noted that Schrödinger also had a 3D version of his equation that he used to find the energy levels of the hydrogen atom. The energy levels were predicted to rousing success agreeing with experimental spectroscopic measurements then. These energy levels had eluded decades of speculative models

If this is conceived as the collective oscillation of the possibly electron-positron (e-p) pairs in a cavity, there are n photons corresponding to energy of $n\hbar\omega_0$ embedded in the collective oscillation. The larger E_n is, the larger the number of photons there is. (However, there is a curious mode at $n = 0$. This corresponds to no photon, and yet, there is a wave function $\Psi_0(q)$. This is the zero-point energy state. This state is there even if the system is at its lowest energy state.)

There were two highlights that were brought about by the wave theory of a particle: the prediction of the energy levels of a hydrogen atom, and the probabilistic interpretation of quantum theory. The meaning of the wave function was not clear to Schrödinger himself. It was Max Born who posit that the wave function was related to the probabilistic interpretation of quantum theory.

The wave function, as shall be seen, implies that the position q of the particle or pendulum is random. Moreover, this position $q(t)$ is mapped to the amplitude $E_x(z,t)$ (for a fixed z) of a traveling field or a cavity field. Hence, it is the amplitude of an electromagnetic oscillation that becomes uncertain and fuzzy as shown in Figure 38.4.

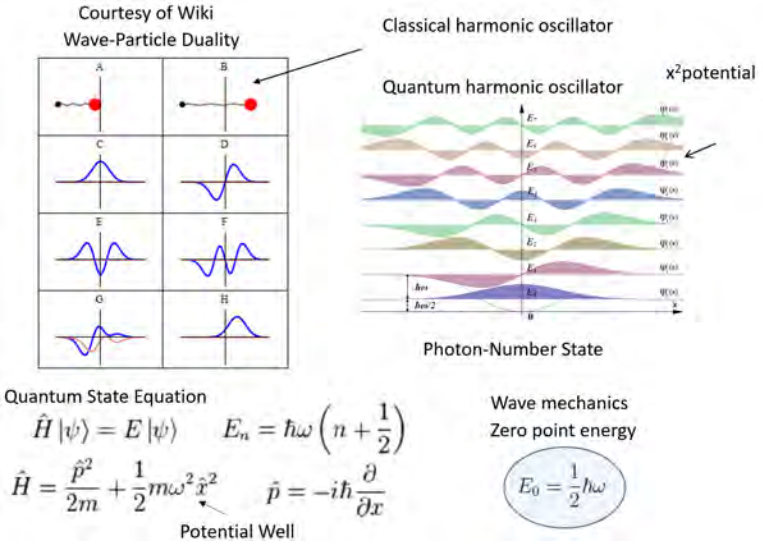


Figure 38.5: Plots of the eigensolutions of the simple quantum pendulum. The photon-number states, $\Psi_n(x)$, are non-classical states because they do not have a classical analogue. To be commensurate with the text of this Chapter, one should replace x in this figure with q. Animation of the left figure are given on the Wiki website. For the detail explanation of various plots, one should consult the wiki page on this matter. Even then, this figure entails lots of information that cannot be digested in one sitting. Alternatively, one should read this chapter repeatedly until the gist of the idea sinks into your soul (courtesy of Wikipedia [320]).

38.4 Representation of Observables in Quantum Theory

Observables are parameters that can be measured in the laboratory, like velocity and position of a pendulum. The most striking feature about quantum theory is that these observables, while they are deterministic variables in the classical world, become random variables in the quantum world. Even though a quantum pendulum is simple, many salient features of quantum theory have emerged. We can study these features in greater detail by elucidating the quantum physics of this very simple quantum system.

38.4.1 Features of Quantum Observables

For a pendulum, the observables are the momentum p and the position q. In the classical world, they are deterministic quantities whose values follow the classical equations of motion. They are also termed conjugate variables: for a dynamic system, they vary in tandem to keep the Hamiltonian a constant for energy conservation.

But in the quantum world, both p and q observables are random variables. Random variables cannot be described by a single number but requires a probability distribution function. In this case, the distribution function can be represented by a vector. In the quantum case, we call this a state vector. Therefore, in the quantum world, the randomness of the observables are best described by the operator-vector pair. In other words, observables p and q are represented by quantum operators $\hat{p}$ and $\hat{q}$. [18] Together with the wave function $\Psi(q,t)$, which can be thought of as a vector, the observables p and q are endowed with random properties. For instance, the observables for the position q and momentum p are now random variables. The random properties of an observable are best described by an operator-vector pairs such as

$$p \Leftrightarrow \{\hat{p}, \Psi(q,t)\}\,, \qquad q \Leftrightarrow \{\hat{q}, \Psi(q,t)\} \tag{38.4.1}$$

In the above, we call operators $\hat{p}$ and $\hat{q}$ the operator representations of the observables p and q. It is important to note that while the operators $\hat{p}$ and $\hat{q}$ are deterministic, even though the observables p and q are random variables.

We shall see next how the quantum observables are endowed with random properties via the association in (38.4.1). For the 1D case, the probability density function for the random variable q is given by $|\Psi(q,t)|^2$, where the probabilty of finding the particle in the interval $[q, q+\Delta q]$ is given by $|\Psi(q,t)|^2\Delta q$. Furthermore, the momentum p also becomes a random variable. It is represented by a momentum operator or differential operator $\hat{p} = -i\hbar\partial/\partial q$. We will discuss how we can find the average values of these quantum observables below.

Dirac Notation and Random Properties of Quantum Observables

At this junction, it is convenient to introduce the Dirac notations[19] In Dirac notation, a vector is denoted as a "ket", written as $|\Psi\rangle$. The conjugate transpose of a vector in Dirac notation is called

[18] In this chapter, we will use "hat" to denote a quantum operator as in many textbooks. These operators, though living in infinite dimensional Hilbert space, are homomorphic to matrix operators in your linear algebra course. Many properties of matrix operators carry over to Hilbert space operators.

[19] It is to be noted that Dirac developed these notation in 1927. His notations are 'homomorphic' to linear algebra or matrix notations which are developed later and are more versatile. But Dirac notations are very popular with physicists.

a "bra" which is denoted as $\langle\Psi_n|$. Hence, the inner product between two vectors is denoted as $\langle\Psi_1|\Psi_2\rangle$ in Dirac notation.[20]

- Now, p and q become random variables, and $|\Psi(q,t)|^2$ is a probability density function. Then

$$\int_{-\infty}^{\infty} dq|\Psi(q,t)|^2 = 1 \Leftrightarrow \underbrace{\langle\Psi(t)|\Psi(t)\rangle}_{\text{Dirac}} = 1 \Leftrightarrow \underbrace{\mathbf{\Psi}(t)^\dagger \cdot \mathbf{\Psi}(t)}_{\text{Matrix}} = 1 \tag{38.4.2}$$

  The above is the normalization condition for all state vectors of a quantum system, since $\Psi(q,t)$ is related to probability density functions. We have chosen to express this normalization condition explicitly using math notation, using Dirac notation, and then using matrix notation.

- The average value or expectation value of the random variable q is now given by

$$\int_{-\infty}^{\infty} dqq|\Psi(q,t)|^2 = \langle q(t)\rangle = \bar{q}(t) \Leftrightarrow \underbrace{\langle\Psi(t)|\hat{q}|\Psi(t)\rangle}_{\text{Dirac}} \Leftrightarrow \underbrace{\mathbf{\Psi}(t)^\dagger \cdot \overline{\mathbf{q}} \cdot \mathbf{\Psi}(t)}_{\text{Matrix}} \tag{38.4.3}$$

  In the above, the operator $\hat{q}$ will be defined later.

- The momentum p is now a random variable related to the differential operator representation $-i\hbar\partial/(\partial q)$. The average or expectation value or average of this random variable p is given by

$$\begin{aligned} -i\hbar \int_{-\infty}^{\infty} dq\Psi^*(q,t)\frac{\partial}{\partial q}\Psi(q,t) &= \langle p(t)\rangle = \bar{p}(t) \\ \Leftrightarrow \underbrace{\langle\Psi(t)|\hat{p}|\Psi(t)\rangle}_{\text{Dirac}} &\Leftrightarrow \underbrace{\mathbf{\Psi}(t)^\dagger \cdot \overline{\mathbf{p}} \cdot \mathbf{\Psi}(t)}_{\text{Matrix}} \end{aligned} \tag{38.4.4}$$

It is to be noted that other properties of a random variable, such as higher moments, can be calculated since the probability density function is known.

38.4.2 More on Quantum Observables

As mentioned before, a quantum observable, which is a random variable, is represented by an operator-vector pair. They endow the observables with random properties [321]. In the above, the operator-vector two-somes, endow the observables with random properties:

$$q \Leftrightarrow \{\hat{q}, |\Psi\rangle\} \qquad p \Leftrightarrow \{\hat{p}, |\Psi\rangle\} \tag{38.4.5}$$

Here, $|\Psi\rangle$ is analogous to a probability density function as defined in (38.4.3) to (38.4.4).

For instance, the averages of the quantum observables, q and p are given by

$$\bar{q} = \langle q\rangle = \langle\Psi|\hat{q}|\Psi\rangle, \qquad \bar{p} = \langle p\rangle = \langle\Psi|\hat{p}|\Psi\rangle \tag{38.4.6}$$

[20] There is a one-to-one correspondence of Dirac notation to matrix algebra notation. $\hat{A}|x\rangle \leftrightarrow \overline{\mathbf{A}} \cdot \mathbf{x}$, $\langle x| \leftrightarrow \mathbf{x}^\dagger$ $\langle x_1|x_2\rangle \leftrightarrow \mathbf{x}_1^\dagger \cdot \mathbf{x}_2$. The preponderance of languages in different communities is like the story of the Tower of Babel.

where we have used Dirac notation for the expectation values of the quantum observables. The formulas for other properties of a random variable can also be computed (see [321]).

In the quantum world, observables are represented by quantum operators (or matrices)-vector pair. By these mere facts, we can conclude about these operator representations.

1. As can be seen above, the expectation value of a quantum observable has to be real valued. This implies that the matrix (or operator) representation of this observable has to be Hermitian.[21]

2. When an observable is measured in the laboratory, the outcome of the measurement of the observable is random (see Figure 38.4). If this random variable is measured many times over and again to obtain its average, this average is related the expectation value of the operator with respect to the probability function described by the state vector $|\Psi\rangle$ given by (38.4.6).

3. The matrix representations of an observable is not unique. Only the expectation value of a quantum observable has to be unique because it corresponds to the average of the measurable random quantity in the laboratory.

4. The matrix (or operator) representations of two different observables are non-commuting, for if they are commutating, they would be diagonalizable by the same set of eigenfunctions. In other words, these observables can be represented by scalar variables instead of matrices or operators. Scalar variables are mutually commuting whereas operators need not be mutually commuting.

The above can be proven easily by those who are mathematically inclined. Or they can be proven using known linear algebra knowledge.

Non-Uniqueness of Operator-Vector Representation

It is to be noted that the above operator-vector representation of the quantum observable is not unique. One can always alter this representation by a unitary transformation. Taking (38.4.4) as an example,

$$\langle p(t)\rangle = \bar{p}(t) = \langle \Psi(t)|\hat{p}|\Psi(t)\rangle = \langle \Psi(t)|\hat{u}\hat{u}^\dagger\hat{p}\hat{u}\hat{u}^\dagger|\Psi(t)\rangle \tag{38.4.7}$$

where $\hat{u}$ is a unitary matrix such that $\hat{u}\hat{u}^\dagger = \hat{I}$. In the above, the definition of $\langle p(t)\rangle = \bar{p}(t)$ remains unchanged, but the operator-vector pair can now be

$$\{\hat{p}', |\Psi'\rangle\} \tag{38.4.8}$$

where $\hat{p}' = \hat{u}^\dagger\hat{p}\hat{u}$, and $|\Psi'\rangle = \hat{u}^\dagger|\Psi\rangle$.

It is interesting to note that when Schrödinger first postulated the operator representation for the momentum, and coordinate observables, he was doing it in coordinate basis, so that $\hat{p} = -i\hbar\partial_q$ and $\hat{q} = \hat{I}q$. These operator representations will be quite different in other basis.

[21] All Hermitian matrix operators have to real eigenvalues.

38.4.3 Quantum Linear Superposition and Quantum Measurements

Before we begin this section, it will be prudent to ask the meaning of a linear system in the quantum world. A quantum system has to be described by a quantum state $|\Psi(t)\rangle$ satisfying the quantum state equation

$$\hat{H}|\Psi(t)\rangle = i\hbar\partial_t|\Psi(t)\rangle \tag{38.4.9}$$

For an energy conserving system, $\hat{H}$ has to be Hermitian and time-independent. Because of this, $\hat{H}$ cannot be a function of $|\Psi(t)\rangle$: This mere fact implies that the above is a linear system. Furthoremore, hermiticity of $\hat{H}$ implies that it has only real eigenvalues. According to linear algebra theory, $\hat{H}$ is also diagonalizable. For subsequent discussion, the solution space of the quantum state equation (38.4.9) will be called the state space.

$$\hat{H} = \sum_{i=1}^{N} \hat{H}_i \tag{38.4.10}$$

We can convert the above equation into an eigenvalue equation by letting $|\Psi_i(t)\rangle = |\Psi_i\rangle e^{-i\omega_i t}$ with $E_i = \hbar\omega_i$. In principle, N can be infinitely large. (The above is also analogous to the separation of variables.)

The principle of linear superposition holds true for a quantum system. The set of eigenvectors $\{|\Psi_i(t)\rangle, i = 1, \ldots, N\}$ forms a complete set that spans the solution space of the quantum state equation.

Since the quantum state equation is linear, the state vector that a quantum system is associated with can be expressed as a linear superposition of state vectors that form a complete set. Mathematically, it is

$$|\Psi(t)\rangle = \sum_{i=1}^{N} a_i|\Psi_i(t)\rangle \tag{38.4.11}$$

Since $|\Psi(t)\rangle$ is related to the probability density function, it is necessary that we have the normalization condition:

$$\langle\Psi(t)|\Psi(t)\rangle = 1 \tag{38.4.12}$$

If $|\Psi_i(t)\rangle$ represents another probability density function of the i-th state of another quantum system, then it is necessary that $\langle\Psi_i|\Psi_i\rangle = 1$. With the normalization condition (38.4.12), then it can be easily shown that

$$|a_1|^2 + |a_2|^2 + |a_3|^2 \ldots + |a_N|^2 = \sum_{i=1}^{N} |a_i|^2 = 1 \tag{38.4.13}$$

Because the $|a_i|^2$ sum up to 1, the above can be given probabilistic interpretation: $|a_i|^2$ is the probability of finding the quantum state to be in state i.

Tenet of Quantum Measurements

The tenet of quantum measurement is one of the "weirdest" things about quantum theory: a quantum state is in a linear superposition of quantum states before the measurement. After the quantum measurement, the quantum state is said to have collapsed to the state that is found by

the measurement. The outcome of the quantum measurement is random, and it collapses to the state that is measured according to the probability distribution $|a_i|^2$.

Quantum Measurements with another Complete Orthonormal Set

In the previous example, we have used the eigenstates of the original quantum system as a way of measuring the quantum state. But this is not necessarily the case. One can also use another complete set of orthonormal states that spans the same space for the measurement space. For instance, one can choose a complete set of eigenfunction $\{|\phi_m\rangle, m = 1, \ldots, N\}$ that spans the same space as the state space to perform the measurement. As the set $\{|\phi_m\rangle\}$ and $\{|\Psi_i\rangle\}$ may not be orthogonal, a projection measurement with $|\phi_m\rangle$ may pick up more than one eigenstates of $|\Psi_i\rangle$. An orthogonal transformation can always be set up to switch the basis from one orthogonal complete set to another orthogonal complete set.

38.5 Beautifying Schrödinger Equation

Rewriting Schrödinger equation as the eigenequation for the photon number state for the quantum harmonic oscillator, we have

$$\hat{H}\Psi_n(q) = \left[-\frac{\hbar^2}{2m}\frac{d^2}{dq^2} + \frac{1}{2}m\omega_0^2 q^2\right]\Psi_n(q) = E_n\Psi_n(q). \tag{38.5.1}$$

where $\hat{H}$ is the Hamiltonian operator, $\Psi_n(q)$ is the eigenfunction also called the photon number state, or Fock state, and E_n is the corresponding eigenvalue. The above can be changed into a dimensionless form first by dividing $\hbar\omega_0$, and then letting $\xi = \sqrt{\frac{m\omega_0}{\hbar}}q$ be a normalized dimensionless variable. The above then becomes

$$\frac{1}{2}\left(-\frac{d^2}{d\xi^2} + \xi^2\right)\Psi_n(\xi) = \frac{E_n}{\hbar\omega_0}\Psi_n(\xi) \tag{38.5.2}$$

(The above is written with a slight abuse of notation, since ξ and q are different variables. In a word, $\Psi(q)$ is not the same as $\Psi(\xi)$.)

Furthermore, the differential equation on the left-hand side in (38.5.2) looks almost like $A^2 - B^2$, and hence motivates its factorization. To this end, we define two beautiful new differential operators

$$\hat{a}^\dagger = \frac{1}{\sqrt{2}}\left(-\frac{d}{d\xi} + \xi\right), \qquad \hat{a} = \frac{1}{\sqrt{2}}\left(\frac{d}{d\xi} + \xi\right) \tag{38.5.3}$$

The first operator above is the creation or raising operator, and the latter operator represents the annihilation or lowering operator.[22] They are also called collectively as ladder operators. The reason for their names is obviated later. With the above definitions of the raising and lowering operators, it is easy to show, by straightforward back substitution, that Schrödinger equation (38.5.2) for a quantum harmonic oscillator can be rewritten more compactly as

$$\frac{1}{2}\left(\hat{a}^\dagger\hat{a} + \hat{a}\hat{a}^\dagger\right)\Psi_n(\xi) = \left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)\Psi_n = \frac{E_n}{\hbar\omega_0}\Psi_n(\xi) \tag{38.5.4}$$

[22] In this course, we assume that quantum operators are 'homomorphic' to matrix operators except that the quantum operators can be infinite dimensional.

Moreover, by more substitution, one can show that the commutator[23]

$$\left[\hat{a}, \hat{a}^\dagger\right] = \hat{a}\hat{a}^\dagger - \hat{a}^\dagger\hat{a} = \hat{I} \tag{38.5.5}$$

This fact has been used to obtain the second equality in (38.5.4).

In mathematics, a function is analogous to a vector. So Ψ_n is the abstract representation of a vector. Consequently, the Hamiltonian operator can now be expressed concisely and abstractly as[24]

$$\hat{H} = \hbar\omega_0 \left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right) \tag{38.5.6}$$

Then we can write (38.5.4) concisely and abstractly as

$$\hat{H}|\Psi_n\rangle = E_n|\Psi_n\rangle \tag{38.5.7}$$

The above is Dirac notation for an operator-vector product, where $|\Psi_n\rangle$ is Dirac's way of indicating an abstract vector. Then the Schrödinger equation (38.3.3) can be written concisely as

$$\hat{H}|\Psi\rangle = i\hbar\partial_t|\Psi\rangle \tag{38.5.8}$$

where we have used Dirac notation $|\Psi\rangle$ to represent a function as a vector which is called a state vector (see Section 36.4). In the above, for energy conservation, $\hat{H}$ is independent of t. Therefore, it can be solved formally to yield[25]

$$|\Psi(t)\rangle = e^{-i\hat{H}t/\hbar}|\Psi(0)\rangle \tag{38.5.9}$$

In abstract Dirac notation, it is

$$\hbar\omega_0 \left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)|\Psi_n\rangle = E_n|\Psi_n\rangle \tag{38.5.10}$$

$$\hat{H}|\Psi_n\rangle = E_n|\Psi_n\rangle \tag{38.5.11}$$

If we denote a photon number state by $\Psi_n(\xi)$ in explicit functional notation, Ψ_n in math notation or $|\Psi_n\rangle$ in Dirac notation, then we have

$$\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)|\Psi_n\rangle = \frac{E_n}{\hbar\omega_0}|\Psi_n\rangle = \left(n + \frac{1}{2}\right)|\Psi_n\rangle \tag{38.5.12}$$

where we have used the fact that $E_n = (n + 1/2)\hbar\omega_0$, since the eigenvalues of this equation is known in closed form. Therefore, by comparing terms in the above, we conclude that

$$\hat{a}^\dagger\hat{a}|\Psi_n\rangle = n|\Psi_n\rangle \tag{38.5.13}$$

[23]The symbol $[\hat{a}, \hat{b}] = \hat{a}\hat{b} - \hat{b}\hat{a}$ is call a commutator.

[24]This is analogous to the abstract linear operator equation $\mathscr{L}f = g$ in the chapter on computational electromagnetics.

[25]This is akin to the state equation in the state-variable approach in control theory.

and the operator $\hat{a}^\dagger\hat{a}$ is also called the number operator because of the above. It is often denoted as

$$\hat{n} = \hat{a}^\dagger\hat{a} \tag{38.5.14}$$

and $|\Psi_n\rangle$ is an eigenvector of $\hat{n} = \hat{a}^\dagger\hat{a}$ operator with eigenvalue n. It can be further shown by direct substitution that[26]

$$\hat{a}|\Psi_n\rangle = \sqrt{n}|\Psi_{n-1}\rangle \qquad \Leftrightarrow \hat{a}|n\rangle \ = \sqrt{n}|n-1\rangle \tag{38.5.15}$$

$$\hat{a}^\dagger|\Psi_n\rangle = \sqrt{n+1}|\Psi_{n+1}\rangle \Leftrightarrow \hat{a}^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle \tag{38.5.16}$$

and hence, their names as lowering and raising operator.[27]

In the above, the Fock state or photon number state in dimensionless unit and coordinate basis is

$$\Psi_n(\xi) = \sqrt{\frac{1}{2^n n!}} e^{-\frac{\xi^2}{2}} H_n(\xi) \tag{38.5.17}$$

where $H_n(\xi)$ is a Hermite polynomial, and the wave function is Gaussian tapered by $e^{-\xi^2/2}$. These Gaussian-tapered Hermite polynomials are orthogonal, since they are the eigenfunctions of an ordinary differential equation, and they can be orthonormalized, namely, that $\langle n|n'\rangle = \delta_{n,n'}$. The ground state, which corresponds to the lowest energy level or smallest eigenvalue, is given by

$$\Psi_0(\xi) = \sqrt{\frac{1}{\pi^{1/4}}} e^{-\xi^2/2} \tag{38.5.18}$$

It is easy to show that $\hat{a}\Psi_0(\xi) = 0$.

38.6 Commutator and Uncertainty Principle

One of the most important consequence of replacing observables with quantum observables with operator representations is the uncertainty principle. In the classical world, observables are scalar variables, and hence, they always commute with each other, namely, $[a, b] = ab - ba = 0$. But observables in the quantum world are represented by matrix operators. These operators need not commute with each other. The physical meaning of noncommuting matrices representing the observables implies that they cannot be simultaneously measured precisely. This is known as the uncertainty principle.

We will summarize this principle as follows:

$$\left[\hat{A}, \hat{B}\right] = i\hat{C} \tag{38.6.1}$$

[26] To do this, it is better to convert these equations in their coordinate space representation. To convert back-forth between these representations, it is better to use the resolution-of-identity procedure [317].

[27] The above notation for a vector could appear cryptic or too terse to the uninitiated. To parse it, one can always down-convert from an abstract notation to a more explicit or down-to-earth notation. Namely, $|n\rangle \to |\Psi_n\rangle \to \Psi_n(\xi)$. One may use the resolution of identity to move between different levels of abstractions. The raising (creation) and lowering (annihilation) operators are the fundamental operators of a quantum harmonic oscillator. The root source of these quantum harmonic operators can be traced to the use of quadratic potential in Schrödinger equation. But they are beautiful operators, and have dominated quantum field theory. [304]

where $\hat{C}$ is Hermitian, then it can be shown that [28, 78] we have

$$\left(\overline{(\Delta A)^2}\right)\left(\overline{(\Delta B)^2}\right) \geq \frac{1}{4}|\langle\hat{C}\rangle|^2 \tag{38.6.2}$$

where the "overline" means "average" with respect to the quantum state vector. The above is the generalized uncertainty principle [28, 322] for two observables A and B. We can take the square root of the above to get

$$\sigma_A \sigma_B \geq \frac{1}{2}|\langle\hat{C}\rangle| \tag{38.6.3}$$

where σ_A and σ_B are the variance of the random observables A and B [78]. We will not show the detail derivations of this principle except for offering its physical interpretation. Interested readers can read up on their derivations in the above references.

38.7 Quantum Information Science and Quantum Interpretation

The random nature of quantum measurements has perplexed researchers at that time. Thus, the most revolutionary revelation of recent years is the interpretation of quantum measurements. There were two prevailing schools of thoughts: the Copenhagen school led by Niels Bohr, and the Einstein school. In the Copenhagen school, it was said that one does not know the quantum state of a system before a measurement. The quantum state can be in a linear superposition of the states. Then the outcome of the measurement of the observable p is random. The average or expected value of this random variable is given in accordance to (38.4.3).But the possible outcomes of measuring the observable p of the system are $\{p_1, p_2, p_3, \ldots, p_n\}$ with probability distributed in accordance to $|a_n|^2$. Upon a measurement, the quantum state of the system "collapses" to the measured state $|P_n\rangle$. Laboratory experiments did confirm that the outcome is random.

However, Einstein, being a *realist* that he was, did not agree with this interpretation. "God does not play dice!", so he said. The above interpretation is too "spooky" to be comfortable with. He posited that a quantum system has already collapsed to the quantum state in a random fashion before the measurement. A measurement only confirms the state to which a quantum system has collapsed to. To explain the randomness of the experimental outcomes, Einstein proposed that there was a hidden random variable involved that decides which state the quantum system was to have collapsed into.

Einstein, being a genius, has his admirers. To prove Einstein right, John Bell came up with an inquality [323]. One can design experiments in the laboratory to measure certain parameters a and b. If Einstein is correct, then laboratory measurements should confirm $a < b$, while if Niels Bohr is correct, then the measurement outcome should confirm that $a > b$. This famous inequality is known as Bell's inequality.

Fortunately, experiments later designed, most notably by Alain Aspect [305, 324], confirmed the Copenhagen school of interpretation as correct, to the dismay of John Bell![28] These were

[28]To me philosophically, this is a beautiful outcome as it implies that our Karma is not written on our forehead when we were born. Our futures are in our own hands:)

important experiments that gave rise to the birth of quantum information science. It means that information can be stored *incognito* in a quantum system before the measurement.

The above linear superposition of states is not possible in our real world, but possible in the quantum world.[29] The linear superposition of states can be choreographed to perform tasks in parallel giving rise to quantum parallelism.

Figure 38.6: Quantum interpretation says that the quantum state of a system is unknown until after a measurement. Only ghosts and angels can do that! (Picture modified from originals by dragoart.com)

38.7.1 Heisenberg Picture versus Schrödinger Picture

Because the quantum state equation can be solved formally in closed form, we can look at the quantum world in two pictures: The Heisenberg picture or the Schrödinger picture. A quantum observable is represented by an operator. Attempt to measure this observable will produce a random result. But if the quantum state of a system, $|\Psi\rangle$, is known, the mean value of the random observable, which is measurable, is given by

$$\bar{O}(t) = \langle \Psi(t)|\hat{O}_s|\Psi(t)\rangle. \tag{38.7.1}$$

But as we said before, the values of $|\Psi\rangle$ and $\hat{O}_s$ are not unique. Alternatively, one can rewrite the above using $|\Psi(t)\rangle = e^{-i\hat{H}t/\hbar}|\Psi(0)\rangle$, which is given formally in (38.5.9). Here, $e^{-i\hat{H}t/\hbar}$ is the time-evolution operator which is a unitary operator. Then

$$\bar{O}(t) = \langle \Psi(0)|e^{i\hat{H}t/\hbar}\hat{O}_s e^{-i\hat{H}t/\hbar}|\Psi(0)\rangle. \tag{38.7.2}$$

In other words, the above can be rewritten as

$$\bar{O}(t) = \langle \Psi(0)|\hat{O}_h(t)|\Psi(0)\rangle. \tag{38.7.3}$$

[29] I call this the "ghost-angel" interpretation of quantum theory.

where

$$\hat{O}_h(t) = e^{i\hat{H}t/\hbar}\hat{O}_s e^{-i\hat{H}t/\hbar}. \tag{38.7.4}$$

The above is a similarity transform with unitary matrix, or a unitary transform. The above illustrates the two different ways to look at the quantum world that produce the same expectation value of a quantum operator. The first way, called the Schrödinger picture, keeps the quantum operator $\hat{O}_s$ to be time independent, but retain the quantum state vector $|\Psi(t)\rangle$ as a function of time. In the second way, called the Heisenberg picture, the operator $\hat{O}_h(t)$ is time dependent, but the state vector $|\Psi(0)\rangle$ is independent of time. Hence, one can derive the equation of motion of an operator in the Heisenberg picture. Also, note that the operators $\hat{O}_h$ and $\hat{O}_s$ in the Heisenberg picture and the Schrödinger picture, respectively, are related to each other by a unitary transform via the time-evolution operator $e^{-i\hat{H}t/\hbar}$.

Equations of Motion of Operators in Heisenberg Picture

Taking the time derivative of an operator in the Heisenberg picture, one concludes that

$$\frac{d\hat{O}_h}{dt} = \frac{i}{\hbar}\left(\hat{H}\hat{O}_h - \hat{O}_h\hat{H}\right) = \frac{i}{\hbar}\left[\hat{H}, \hat{O}_h\right] \tag{38.7.5}$$

where the commutator $\left[\hat{H}, \hat{O}_h\right] = \hat{H}\hat{O}_h - \hat{O}_h\hat{H}$ is used to abbreviate the above equation. The above is the Heisenberg equation of motion for quantum operators. One can apply the above to the operators $\hat{p}$, $\hat{q}$, $\hat{a}$ and $\hat{a}^\dagger$ operators in Heisenberg picture to arrive at their equations of motion. For instance,

$$\frac{d\hat{p}}{dt} = \frac{i}{\hbar}\left[\hat{H}, \hat{p}\right], \qquad \frac{d\hat{q}}{dt} = \frac{i}{\hbar}\left[\hat{H}, \hat{q}\right] \tag{38.7.6}$$

By repeated use of the commutator $[\hat{p}, \hat{q}] = -i\hbar\hat{I}$, we can show that [322, 316]

$$\frac{d\hat{p}}{dt} = -\frac{\partial\hat{H}(\hat{p}, \hat{q})}{\partial\hat{q}}, \qquad \frac{d\hat{q}}{dt} = \frac{\partial\hat{H}(\hat{p}, \hat{q})}{\partial\hat{p}} \tag{38.7.7}$$

They can also be derived using energy-conservation arguments [316]. They are 'homomorphic' to the classical Hamilton equations. Thus we call them the quantum Hamilton equations.

The above shows that the equations of motion in the Heisenberg picture is very similar to the classical equations of motion. It also greatly simplifies the derivation of quantum equations of motion in the quantum world. Many classical equations of motion are derivable using classical Hamiltonian theory. As we shall see, the use of 'homomorphism' greatly simplifies the derivation of quantum equations of motion from quantum Hamiltonian theory [321].

Exercises for Lecture 38

Problem 38-1: This problem refers to Chapter 38. If you are busy, you can pick four of seven in the following parts.

(i) Who first postulated that the energy of electromagnetic field is quantized into packets given by the formula $\hbar\omega$? Find this packets of energy for a 10 GHz microwave photon, and an optical photon of 500 nm in wavelength.

(ii) How did Schrödinger arrive at the motivation for Schrödinger equation (38.3.3)?

(iii) In classical theory, the position and momentum of a particle is given by q and p. But in quantum theory, the position and momentum of a particle are random variables and become improbable. How are these random properties expressed in quantum theory?

(iv) Show that when the system is in the photon-number state or Fock state, the expectation value of the position operator $\hat{q}$ and the momentum operator $\hat{p}$ are zero.

(v) Explain why two non-commutating operators cannot share a common eigenvector, and how this is related to the uncertainty principle.

(vi) Derive (38.7.5), the equation of motion for a quantum operator.

(vii) Derive (38.7.7).

Problem 38-2: Verify the mathematical itemized assertions in Section 38.4.2 of the text. The proof can be rather simple for the math-inclined, and otherwise, they can be found in many textbooks.

Problem 38-3: Show that the higher moments of an observable, which is a random variable, are the same in both the Heisenberg picture and the Schrödinger picture.

Chapter 39

Quantum Coherent State of Light and More

As mentioned in the previous lecture, the discovery of Schrödinger wave equation was a rousing success! When it was discovered by Schrödinger, and applied to the very simple hydrogen atom, its eigensolutions, especially the eigenvalues E_n coincided beautifully with spectroscopy experiments of the hydrogen atom. Since the electron wave functions inside a hydrogen atom does not have a classical analog, less was known about their physical meanings.

But in QED (quantum electrodynamics) and QO (quantum optics), these wavefunctions have to be connected with classical electromagnetic oscillations. As seen previously, classical electromagnetic oscillations resemble those of a classical pendulum. In the quantum world, the original eigenstates of the quantum pendulum were the photon number states also called the Fock states. These quantum wave oscillations, the Fock states, do not resemble the classical oscillations of a classical pendulum at all! Their connection to the classical pendulum was tenuous, but required by the correspondence principle—quantum wave phenomena should resemble classical phenomena in the high energy (high frequency, short wavelength) limit.[1] This connection was finally established by using the coherent state.[2]

After connecting the classical pendulum to the quantum pendulum using the coherent state, we will study the use of quantum electromagnetics/optics for a communication problem as an illustration of this knowledge. To do this, we have to connect classical electromagnetics to the Hamiltonian theory, and then to quantum electromagnetics. With this, we are better able to understand this quantum knowledge.

[1] In its simplest and approximate form, Schrödinger equation can be described by $\frac{(\hbar k)^2}{2m} + V = E = \hbar\omega$. In the high frequency ω limit, E has to increase, and the value of k^2 has to increase to balance this equation. Hence, in the high frequency limit of a quantum pendulum, the associated quantum k has to increase, implying shorter wavelengths.

[2] Quantum electrodynamics is also the precursor to quantum field theory [45].

39.1 The Quantum Coherent State

We have seen that the photon number states of a quantum pendulum do not have a classical correspondence as the average or expectation values of the position and momentum of the pendulum in the photon-number state (often just called a number state) are always zero for all time for this state. Hence, photon-number states are also known as non-classical states.[3]

Therefore, we have to seek a time-dependent quantum state that has the classical equivalence of a pendulum. This is offered by the coherent state, which is the contribution of many researchers, most notably, Roy Glauber (1925–2018) [325] in 1963, and George Sudarshan (1931–2018) [326]. Glauber was awarded the Nobel prize in 2005.

We like to emphasize again that the mode of an electromagnetic field inside a cavity, or a traveling wave oscillation in an open space, is 'homomorphic' to the oscillation of a classical pendulum if we gaze at it for a fixed location. Hence, we first connect the oscillation of a quantum pendulum to that of a classical pendulum. Then we can re-connect the oscillation of a classical electromagnetic mode to that of a quantum electromagnetic mode. In other words, we re-connect the classical pendulum to the quantum pendulum. The coherent state is a linear superposition of photon number states that makes it look like a localized wave packet. As such, a coherent state can make a quantum pendulum resemble a classical pendulum in the correspondence-principle limit [317].

The above discussion is akin to the fact that a delta pulse $\delta(z - ct)$ is a solution to the wave equation (see Section 3.2). The delta pulse can also be expanded in terms of superposition of plane waves $\exp(-i\omega t + ikz)$ with $k = \omega/c$. But the plane wave modes are not at all localized, whereas the delta pulse is. By the same token, the photon number states are not localized, but a linear superposition of them forms the coherent state that we shall show to correspond to a localized pulse [317].

Another important point to note is that a photon is a very different quantum particle from an electron. A photon is a boson, whereas an electron is a fermion. Bosons like to behave cooperatively with each other, moving in unison through space. But fermions have to obey the Pauli's exclusion principle, meaning that they have to be opposite in quantum states (one spin up and the other spin down) before they can pair up with each other. The famous example is the formation of Cooper pairs between two electrons of opposite spins. This happens when the temperature is low enough. As a consequence, Cooper pairs behave like bosons, allowing them to move cooperatively and seamlessly through a lattice if the thermal vibration of the lattice is low enough. This was first explained by the famous BCS (Bardeen, Cooper, Schrieffer) theory [327] and also explained in textbooks such as [73].

A photon is a wonderful particle: First, it is massless; and second, it zips around at the speed of light. Third, it has been known to traverse galaxies that are billions of light years away. Photons have been known to propagate cooperatively through an optical fiber a with loss of about 0.25 dB/km [328]. Photons have been used for quantum teleportation [329] [4] and they have been used to confirm the Bell's inequality [305]. Recently, quantum parallelism has been demonstrated by using a boson sampler in the manner of the simple JiuZhang quantum computer [330].

[3] A photon number state to a quantum state is like a Fourier mode to a classical signal. It alone does not have a physical meaning because it has infinite duration. A physical signal has to be of finite duration *de rigueur*. But a summation of Fourier modes can be used to reconstitute any signals of finite duration. A single photon number state does not resemble a physical wave function of a quantum state. But a linear superposition of the photon number state can reconstitute a physical quantum wave function that is physical with a finite duration!

[4] A simple intro to quantum teleportation can be found in [78].

[329]

As we have seen in the previous chapter, electromagnetic fields have to be elevated into quantum fields to describe the physics of photons correctly. It is on top of this quantum fields that a photon "rides" on.[5]

39.2 Some Details on the Coherent States

As one cannot see the characteristics of a classical pendulum emerging from the photon number states, one needs another way of bridging the quantum world with the classical world as required by the correspondence principle. We need to find a wave-packet description of the quantum pendulum that will bridge this gap. This is the role of the coherent state: It will reveal the correspondence principle, with a classical pendulum emerging from a quantum pendulum when the energy of the pendulum is large.

The derivation of the coherent state is more math than physics. So we will describe it in the next subsection below, but its derivation can be found in many textbooks [331][p. 44], [332][p. 158]. To say succinctly, the coherent state is the eigenstate of the annihilation operator, namely that

$$\hat{a}|\alpha\rangle = \alpha|\alpha\rangle \tag{39.2.1}$$

Here, we use α as an eigenvalue as well as an index or identifier of the state $|\alpha\rangle$.[6] Since the number state $|n\rangle$ is complete,[7] the coherent state $|\alpha\rangle$ can be expanded in terms of the number state $|n\rangle$.[8]

The coherent state can be derived to be [332, 331]

$$|\alpha\rangle = e^{-|\alpha|^2/2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle \tag{39.2.2}$$

(Also since $\hat{a}$ is a non-Hermitian operator, its eigenvalue α can be a complex number.) The above can be proved easily by back substitution. A closed form expression also exists for the above summation [333, p. 50], namely, in coordinate basis,

$$\Psi_\alpha(\xi) = \langle \xi|\alpha\rangle = \left(\frac{\omega}{\pi\hbar}\right)^{1/4} e^{-|\alpha|^2/2} e^{\xi^2/2} e^{-(\xi-\alpha/\sqrt{2})^2} \tag{39.2.3}$$

(The coherent state can also be expressed as a displaced harmonic oscillator [73], which has a different Hamiltonian from a simple harmonic oscillator.)

[5] Also, quantum electrodynamics gave birth to quantum field theory [45, 304].

[6] This notation is cryptic and terse, but one can always down-convert it as $|\alpha\rangle \to |f_\alpha\rangle \to f_\alpha(\xi)$ to get a more explicit notation with an intuitive feel. However, using eigenvalue to index a quantum state is a potential source of confusion. For instance, when we denote a photon number state by $|n\rangle$, n is not the eigenvalue of the Schrödinger Hamiltonian but the eigenvalue of the number operator $\hat{n}$. One has to be wary of such confusion when parsing these notations in the physics community.

[7] Because it is the eigenstate of a Hermitian operator, the number operator $\hat{n} = \hat{a}^\dagger \hat{a}$.

[8] The math-physics of the problem comes together beautifully: the photon number state (also called the Fock state) is the eigenstate of the number operator $\hat{n} = \hat{a}^\dagger \hat{a}$, but the coherent state is the eigenstate of the annihilation operator $\hat{a}$. But these two operators, the number state $\hat{n}$ and annihilation operator $\hat{a}$ do not commute. Therefore, they do not share the same set of eigenstates or eigenfunctions. As such, these two operators cannot be diagonalized by a common set of eigenfunctions.

Since the coherent state is a linear superposition of the photon number states, an average number of photons can be associated with the coherent state. If the average number of photons embedded in a coherent state is N, then it can be shown that $N = |\alpha|^2$. As shall be shown, α is related to the amplitude of the quantum oscillation: The more photons there are in a coherent state, the larger $|\alpha|$ is.

Derivation of the Coherent States

Since the number state, which is the eigenstate of the number operator $\hat{n}$, $|n\rangle$ is complete, the coherent state $|\alpha\rangle$ can be expanded in terms of the number state $|n\rangle$. Or that

$$|\alpha\rangle = \sum_{n=0}^{\infty} C_n |n\rangle \tag{39.2.4}$$

When the annihilation operator is applied to the above, we have

$$\hat{a}|\alpha\rangle = \sum_{n=0}^{\infty} C_n \hat{a}|n\rangle = \sum_{n=1}^{\infty} C_n \hat{a}|n\rangle = \sum_{n=1}^{\infty} C_n \sqrt{n}|n-1\rangle = \sum_{n=0}^{\infty} C_{n+1}\sqrt{n+1}|n\rangle \tag{39.2.5}$$

The last equality follows from changing the variable of summation from n to $n+1$. Equating the above with $\alpha|\alpha\rangle$ on the right-hand side of (39.2.1), then

$$\sum_{n=0}^{\infty} C_{n+1}\sqrt{n+1}|n\rangle = \alpha \sum_{n=0}^{\infty} C_n |n\rangle \tag{39.2.6}$$

By the orthonormality of the number states $|n\rangle$ and the completeness of the set, we arrive at

$$C_{n+1} = \alpha C_n / \sqrt{n+1} \tag{39.2.7}$$

Or recursively

$$C_n = C_{n-1}\alpha/\sqrt{n} = C_{n-2}\alpha^2/\sqrt{n(n-1)} = \ldots = C_0 \alpha^n/\sqrt{n!} \tag{39.2.8}$$

Consequently, the coherent state $|\alpha\rangle$ is

$$|\alpha\rangle = C_0 \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}}|n\rangle \tag{39.2.9}$$

But due to the probabilistic interpretation of a quantum state, the state vector $|\alpha\rangle$ is normalized to one, or that[9]

$$\langle\alpha|\alpha\rangle = 1 \tag{39.2.10}$$

[9]The expression can be written more explicitly as $\langle\alpha|\alpha\rangle = \langle f_\alpha | f_\alpha\rangle = \int_{-\infty}^{\infty} d\xi f_\alpha^*(\xi) f_\alpha(\xi) = 1$.

Then

$$\begin{aligned}\langle\alpha|\alpha\rangle &= C_0^* C_0 \sum_{n,n'}^{\infty} \frac{\alpha^n}{\sqrt{n!}} \frac{\alpha^{n'}}{\sqrt{n'!}} \langle n'|n\rangle \\ &= |C_0|^2 \sum_{n=0}^{\infty} \frac{|\alpha|^{2n}}{n!} = |C_0|^2 e^{|\alpha|^2} = 1 \end{aligned} \tag{39.2.11}$$

Therefore, $C_0 = e^{-|\alpha|^2/2}$ for normalization, or that the expression for the coherent state as given before in (39.2.2)

$$|\alpha\rangle = e^{-|\alpha|^2/2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle \tag{39.2.12}$$

In the above, to reduce the double summations into a single summation, we have made use of that $\langle n'|n\rangle = \delta_{n'n}$, or that the photon-number states are orthonormal. Also since $\hat{a}$ is not a Hermitian operator, its eigenvalue α can be a complex number.[10]

39.2.1 Time Evolution of a Quantum State

As mentioned before, the Schrödinger equation for the quantum state (also called the state vector) of a quantum particle can be written concisely as

$$\hat{H}|\Psi\rangle = i\hbar\partial_t|\Psi\rangle \tag{39.2.13}$$

The above not only entails the form of Schrödinger equation, it is the form of the general quantum state equation. Since $\hat{H}$ is time independent, the formal solution to the above is[11]

$$|\Psi(t)\rangle = e^{-i\hat{H}t/\hbar}|\Psi(0)\rangle \tag{39.2.14}$$

We remind ourselves that the Hamiltonian $\hat{H}$ of a quantum pendulum is of the form $\hbar\omega_0\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right) = \hbar\omega_0\left(\hat{n} + \frac{1}{2}\right)$ where $\hat{n} = \hat{a}^\dagger\hat{a}$. Here, $\hat{n}$ is Hermitian, and its eigenvector is $|n\rangle$ with eigenvalue n. One can apply this to the photon number state with $\hat{H}$ being that of the quantum pendulum and that $|n\rangle$ is the eigenvector or eigenstate of $\hat{H}$. Then it is quite easy to show that

$$\hat{H}|n\rangle = \hbar\omega_0\left(n + \frac{1}{2}\right)|n\rangle = \hbar\Omega_n|n\rangle \tag{39.2.15}$$

In other words, the eigenvalue of the Hamiltonian is $\hbar\Omega_n = \hbar\omega_0\left(n + \frac{1}{2}\right)$. Then

$$e^{-i\hat{H}t/\hbar}|n\rangle = e^{-i\Omega_n t}|n\rangle \tag{39.2.16}$$

where $\Omega_n = \left(n + \frac{1}{2}\right)\omega_0$.

[10] The eigenstates of a Hermitian operator can be proven to be complete, but that of a non-Hermitian operator has eigenstates that need not be complete until proven so. This fact has been used to study solutions of iterative method with great success in Trefethan [334].

[11] The following is a function of an operator acting on a vector, for a review, one can read the Appendix.

Time Evolution of the Coherent State

Using the above time-evolution operator, then the time dependent coherent state, after using (39.2.2), evolves in time as[12]

$$|\alpha,t\rangle = e^{-i\hat{H}t/\hbar}|\alpha\rangle = e^{-|\alpha|^2/2}\sum_{n=0}^{\infty}\frac{\alpha^n e^{-i\hat{H}t/\hbar}}{\sqrt{n!}}|n\rangle = e^{-|\alpha|^2/2}\sum_{n=0}^{\infty}\frac{\alpha^n e^{-i\Omega_n t}}{\sqrt{n!}}|n\rangle \tag{39.2.17}$$

In the above, $|\alpha,t\rangle$ is time dependent, and hence, it is in the Schrödinger picture. The state $|\alpha\rangle$ can be thought of as the initial value of $|\alpha,t\rangle$ at $t=0$. Also, $|n\rangle$ on the right-hand side are independent of time, or evaluated at $t=0$. The last equality above is established with the help of (39.2.16). By letting $\Omega_n = \omega_0\left(n+\frac{1}{2}\right)$, the above can be rewritten as

$$|\alpha,t\rangle = e^{-i\omega_0 t/2}e^{-|\alpha|^2/2}\sum_{n=0}^{\infty}\frac{\left(\alpha e^{-i\omega_0 t}\right)^n}{\sqrt{n!}}|n\rangle \tag{39.2.18}$$

By rewriting $|\alpha,t\rangle$ as such, it is clearly an eigenvector of the annihilation operator $\hat{a}$ because it is of the same form as in (39.2.2), except for a multiplicative factor $e^{-i\omega_0 t/2}$, and with the substitution $\alpha \to \alpha e^{-i\omega_0 t}$.

Now we see that the last factor in (39.2.18) is similar to the expression for a coherent state in (39.2.2) with $\alpha \to \alpha e^{-i\omega_0 t}$. Therefore, we can express the above more succinctly by replacing α in (39.2.2) with $\tilde{\alpha} = \alpha e^{-i\omega_0 t}$ as

$$|\alpha,t\rangle = e^{-i\omega_0 t/2}|\alpha e^{-i\omega_0 t}\rangle = e^{-i\omega_0 t/2}|\tilde{\alpha}\rangle \tag{39.2.19}$$

The above is clearly an eigenvector of the annihilation operator $\hat{a}$. Hence, we will use the eigenvalue $\tilde{\alpha}$ to index this eigenvector $|\tilde{\alpha}\rangle$ but $\tilde{\alpha} = \alpha e^{-i\omega_0 t}$ is a complex number which is also a function of time t. It is to be noted that in the coherent state in (39.2.18), the photon number states time-evolve coherently together in a manner to result in a phase shift $e^{-i\omega_0 t}$ in the eigenvalue giving rise to a new eigenvalue $\tilde{\alpha}$!

39.3 More on the Creation and Annihilation Operator

As seen in the photon-number states (also called Fock state), it is not apparent that the classical oscillation of the pendulum will emerge from the photon number state, which is the solution of Schrödinger equation. This is required by the correspondence principle. Hence it is prudent to see if this physical phenomenon emerges with the coherent state. In order to connect the quantum pendulum to a classical pendulum via the coherent state, we will introduce some new operators.

[12]Note that $|\alpha,t\rangle$ is an abstract shorthand for $f_\alpha(\xi,t)$.

Since the creation and annihilation operators in coordinate basis are[13]

$$\hat{a}^\dagger = \frac{1}{\sqrt{2}}\left(-\frac{d}{d\xi} + \xi\right) \tag{39.3.1}$$

$$\hat{a} = \frac{1}{\sqrt{2}}\left(\frac{d}{d\xi} + \xi\right) \tag{39.3.2}$$

We can relate $\hat{a}^\dagger$ and $\hat{a}$, which are non-Hermitian, to the normalized momentum operator $\hat{\pi}$ and the normalized position operator $\hat{\xi}$, previously defined, which are Hermitian. They have to be Hermitian since they are the operator representations of the normalized quantum observables, π and ξ which are real valued.

Then from the above,

$$\hat{a}^\dagger = \frac{1}{\sqrt{2}}\left(-i\hat{\pi} + \hat{\xi}\right) \tag{39.3.3}$$

$$\hat{a} = \frac{1}{\sqrt{2}}\left(i\hat{\pi} + \hat{\xi}\right) \tag{39.3.4}$$

From the above, by subtracting and adding the two equations, we arrive at new operators in coordinate basis as

$$\hat{\xi} = \frac{1}{\sqrt{2}}\left(\hat{a}^\dagger + \hat{a}\right) = \xi\hat{I} \tag{39.3.5}$$

$$\hat{\pi} = \frac{i}{\sqrt{2}}\left(\hat{a}^\dagger - \hat{a}\right) = -i\frac{d}{d\xi} \tag{39.3.6}$$

The above algebra is similar to that for finding the real and imaginary part of a complex number. Also, notice that $\hat{\xi}$ and $\hat{\pi}$ are Hermitian operators whereas the original ladder operators $\hat{a}$ and $\hat{a}^\dagger$ are not. We can think of $\hat{a}$ and $\hat{a}^\dagger$ are operator representation of the complex amplitudes a and a^*.

It is to be noted that $\hat{a}$ and $\hat{a}^\dagger$ time evolve as $e^{-i\omega_0 t}$ and $e^{i\omega_0 t}$ respectively. In a word, they are rotating or counter rotating waves. By decomposing them in terms of $\hat{\pi}$ and $\hat{\xi}$, we are decomposing them in terms of in-phase and quadrature components just as in traditional electrical engineering!

Furthermore, the expectation value of the creation and annihilation operators with respect to the photon number state is zero. For example,

$$\langle n|\hat{a}|n\rangle = 0 \tag{39.3.7}$$

because $\hat{a}|n\rangle$ always lowers the photon number state by one, and that $\langle n|n-1\rangle = 0$ always due their orthogonality. Taking the conjugate transpose of the above, we always get

$$\langle n|\hat{a}^\dagger|n\rangle = 0 \tag{39.3.8}$$

From this, we gather that the expectation values of $\hat{\xi}$ and $\hat{\pi}$ with respect to the photon number state is always zero. That is why the number state (photon number state) is a non-classical state.

[13]This is not entirely koshered as the left-hand side is an abstract operator but the right-hand side is in coordinate basis. There is a formal way to go between abstract operator and its coordinate basis representation as shown in the Appendix under Resolution of Identity (see also [317]).

39.3.1 The Correspondence Principle for a Pendulum

From the above, it is seen that the expectation values of ξ and π have to be taken with respect to another state to see the correspondence principle, viz., they should look like the position and momentum of a classical pnedulum. This can be found by taking their expectation values with respect to the coherent state $|\alpha\rangle$. We begin with the observable for normalized position ξ or that

$$\langle\alpha|\hat{\xi}|\alpha\rangle = \frac{1}{\sqrt{2}}\langle\alpha|\hat{a}^\dagger + \hat{a}|\alpha\rangle \tag{39.3.9}$$

We write the expectation value in the above form, as will be obvious later, to exploit the fact that the coherent state $|\alpha\rangle$ is an eigenstate of the annihilation operator $\hat{a}$. To this end, we take the complex conjugate transpose of (39.2.1)[14], we have

$$\langle\alpha|\hat{a}^\dagger = \langle\alpha|\alpha^* \tag{39.3.10}$$

and (39.3.9) becomes

$$\bar{\xi} = \langle\xi\rangle = \langle\alpha|\hat{\xi}|\alpha\rangle = \frac{1}{\sqrt{2}}\left(\alpha^* + \alpha\right)\underbrace{\langle\alpha|\alpha\rangle}_{=1} = \sqrt{2}\Re e\left(\alpha\right) \neq 0 \tag{39.3.11}$$

Repeating the exercise for time-dependent case, when we let $\alpha \to \tilde{\alpha}(t) = \alpha e^{-i\omega_0 t}$, then, letting $\alpha = |\alpha|e^{-i\Psi}$ or $\tilde{\alpha}(t) = |\alpha|e^{-i\Psi - i\omega_0 t}$ yields

$$\bar{\xi}(t) = \langle\xi(t)\rangle = \langle\tilde{\alpha}(t)|\hat{\xi}|\tilde{\alpha}(t)\rangle = \frac{1}{\sqrt{2}}\left[\tilde{\alpha}^*(t) + \tilde{\alpha}(t)\right]\underbrace{\langle\tilde{\alpha}(t)|\tilde{\alpha}(t)\rangle}_{=1} = \sqrt{2}\Re e\left(\tilde{\alpha}(t)\right) \neq 0 \tag{39.3.12}$$

In both (39.3.11) and (39.3.12), we have made use of that $\langle\alpha|\alpha\rangle = 1$, and $\langle\tilde{\alpha}(t)|\tilde{\alpha}(t)\rangle = 1$ which are the normalization conditions for a quantum states. Then, letting $\tilde{\alpha}(t) = \alpha e^{-i\omega_0 t}$ where $\alpha = |\alpha|e^{-i\Psi}$ is also a complex number, we have

$$\bar{\xi}(t) = \langle\xi(t)\rangle = \sqrt{2}|\alpha|\cos(\omega_0 t + \Psi) \tag{39.3.13}$$

In the above, we use $\xi(t)$ to denote the random variable. So $\langle\xi(t)\rangle$ refers to the average of the random variable $\xi(t)$, or $\bar{\xi}(t)$ that is also a function of time.

By the same token,

$$\bar{\pi} = \langle\pi\rangle = \langle\alpha|\hat{\pi}|\alpha\rangle = \frac{i}{\sqrt{2}}\left(\alpha^* - \alpha\right)\langle\alpha|\alpha\rangle = \sqrt{2}\Im m\left(\alpha\right) \neq 0 \tag{39.3.14}$$

For the time-dependent case, we let $\alpha \to \tilde{\alpha}(t) = \alpha e^{-i\omega_0 t}$ to arrive at

$$\bar{\pi}(t) = \langle\pi(t)\rangle = -\sqrt{2}|\alpha|\sin(\omega_0 t + \Psi) \tag{39.3.15}$$

Hence, we see that the expectation values of the normalized coordinate and momentum just behave like a classical pendulum. There is however a marked difference: These values, ξ and π, which

[14] Dirac notation is homomorphic with matrix algebra notation. $(\overline{\mathbf{a}}\cdot\mathbf{x})^\dagger = \mathbf{x}^\dagger\cdot(\overline{\mathbf{a}})^\dagger$.

are quantum observables in the quantum world, are random variables: they have mean values with standard deviations or variances that are non-zero.Thus, they have quantum "fluctuation" or quantum noise associated with them. Since the quantum pendulum is homomorphic with the oscillation of a quantum electromagnetic mode, the amplitude of a quantum electromagnetic mode will have a mean and a "fluctuation" as well. Now, these are quantum noise associated with a quantum observable. However, this does not imply that these observables are fluctuating with respect to time: they are subject to the probablistic interpretation of quantum theory!

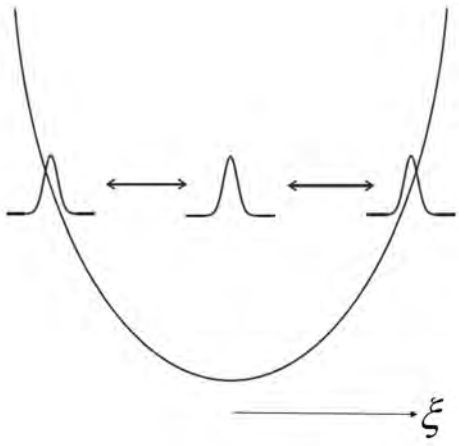


Fig. 3.4. A coherent-state wave function moves through the harmonic oscillator potential, between the classical turning points, without dispersion.

Figure 39.1: The time evolution of the coherent state. It is a wave packet that follows the motion of a classical pendulum or harmonic oscillator (courtesy of Gerry and Knight [333]).

.

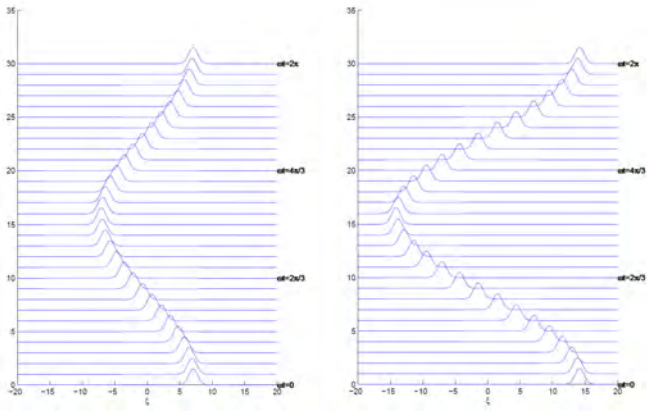


Figure 39.2: The time evolution of the coherent state for different α's. The left figure is for $\alpha = 5$ while the right figure is for $\alpha = 10$. Recall that $N = |\alpha|^2$. Again, it shows the motion of a wave packet.

.

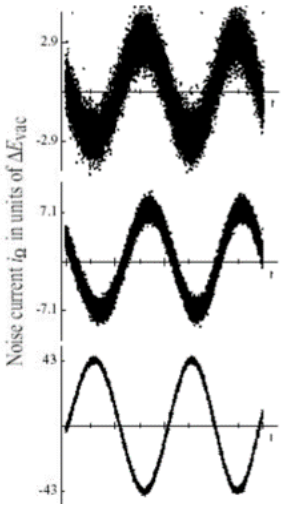


Figure 39.3: The time evolution of the coherent state at the very low signal level. It is sinusoidal, but with great uncertainty innate in it (courtesy of Wikipedia [335]).

.

39.3.2 Mean and Variance of the Amplitude of the Coherent State

The great attribute about the coherent state is that the expectation value or the average of the field amplitude squared approaches the classical limit as the number of photons embedded in this state approaches infinity. It is its variance of its amplitude that distinguishes the coherent state from the classical field.

Considering the case of the quantum pendulum, we can represent the complex amplitude of the oscillation by an operator $\hat{a}$. As we have shown previously, the Hamiltonian of such a quantum system is given by

$$\hat{H} = \frac{1}{2}\hbar\omega_0 \left(\hat{a}\hat{a}^\dagger + \hat{a}^\dagger\hat{a}\right) \tag{39.3.16}$$

In the above, $\hat{a}$ and $\hat{a}^\dagger$ are functions of time in the Heisenberg picture. One can easily show that [331, eq. 2.20]

$$\hat{a}(t) = \hat{a}_0 e^{-i\omega t}, \quad \hat{a}^\dagger(t) = \hat{a}_0^\dagger e^{+i\omega t} \tag{39.3.17}$$

It is interesting to note that $\hat{a}$ and $\hat{a}^\dagger$ are functions of time, in the Heisenberg picture, but when they are back-substituted into the Hamiltonian, the Hamiltonian remains to be time independent, which is required by energy conservation.

It is important to note that the lowering and raising operators are themselves not representations of real-value observables. However, the "real" part and the "imaginary" part of these operators are representations of quantum observables because they are Hermitian.

Another quantum observable is the photon number n, which is real value, and its operator representation is Hermitian. Such is the number operator

$$\hat{n} = \hat{a}^\dagger\hat{a} \tag{39.3.18}$$

The above is clearly Hermitian. Moreover, its expectation value with respect to the coherent state is easily found to be real-valued

$$\langle\alpha|\hat{n}|\alpha\rangle = |\alpha|^2 = \bar{n} \tag{39.3.19}$$

where $\bar{n}$ means the average value of n. We can also easily find the expectation value of the $\hat{n}^2 = \hat{a}^\dagger\hat{a}\hat{a}^\dagger\hat{a}$ operator to be

$$\langle\hat{n}^2\rangle = \langle\alpha|\hat{a}^\dagger\hat{a}\hat{a}^\dagger\hat{a}|\alpha\rangle = \langle\alpha|\hat{a}^\dagger[1 + \hat{a}^\dagger\hat{a}]|\alpha\rangle \tag{39.3.20}$$

In the above, we have used the commutator relation that $[\hat{a}\hat{a}^\dagger - \hat{a}^\dagger\hat{a}] = 1$ to expand the term. The above can be easily shown to be $|\alpha|^2 + |\alpha|^4$.

Notice that in the above, n is a quantum observable which is random and hence, its matrix representation $\hat{n}$ is Hermitian. From the above, we can derive the variance of the quantum observable n, a random variable, or the photon number of the coherent state which is

$$\sigma_n^2 = \langle\hat{n}^2\rangle - \bar{n}^2 = |\alpha|^2 \tag{39.3.21}$$

If the standard deviation is divided by the average of the signal, one has

$$\frac{\sigma_n}{\bar{n}} = \frac{1}{\alpha} \tag{39.3.22}$$

The above implies that the larger the signal, the smaller is the noise (see Figure 39.3).

39.4 Connecting Quantum Pendulum to Electromagnetic Oscillator and its Hamiltonian

In this section, we will connect classical electromagnetics with Hamiltonian theory. Then it is easier to connect classical electromagnetics to quantum electromagnetics. With quantum electromagnetics, we can study how a photon rides on a quantum wave field. Then we can use this quantum wave field for quantum communications.

If we fix our gaze at the field of an electromagnetic oscillator, we see that the electromagnetic oscillator is similar or 'homomorphic' to a pendulum. This electromagnetic oscillator can be a cavity mode, or a traveling electromagnetic wave observed at a fixed location. The classical Hamiltonian for a pendulum is

$$H = T + V = \frac{p^2}{2m} + \frac{1}{2}m\omega_0^2 q^2 = E \tag{39.4.1}$$

where E is the total energy of the system. In the above, ω_0 is the resonant frequency of the classical pendulum.

Each electromagnetic mode or oscillator is 'homomorphic' to the simple pendulum. In order for each electromagnetic mode to be homomorphic the simple pendulum, we have to replace ω_0, the resonant frequency of the pendulum, with ω_l, the resonant frequency of the l-th electromagnetic mode or oscillator. Or each cavity mode or a traveling wave with resonant frequency ω_l is homomorphic to a pendulum with resonant frequency ω_0. We have also shown that when the classical pendulum is elevated to be a quantum pendulum, then $H \to \hat{H}$, where $\hat{H} = \hbar\omega_l \left(\hat{a}^\dagger \hat{a} + \frac{1}{2}\right)$. Then the corresponding Schrödinger equation becomes

$$\hat{H}|\Psi(t)\rangle = \hbar\omega_l \left(\hat{a}^\dagger \hat{a} + \frac{1}{2}\right)|\Psi(t)\rangle = i\hbar\partial_t|\Psi(t)\rangle \tag{39.4.2}$$

Our next task is to connect this quantum pendulum to the electromagnetic oscillator.

39.4.1 Semi-Classical Picture of a Plane Wave

In general, the total energy or the Hamiltonian of an electromagnetic system is

$$H = \frac{1}{2}\int_V d\mathbf{r}\left[\varepsilon \mathbf{E}^2(\mathbf{r},t) + \frac{\mathbf{B}^2(\mathbf{r},t)}{\mu}\right]. \tag{39.4.3}$$

It is customary to write this Hamiltonian in terms of scalar and vector potentials. We assume the Coulomb gauge $\nabla \cdot \mathbf{A} = 0$ with $\Phi = 0$.[15] When $\Phi = 0$, the Coulomb gauge and the Lorenz gauge are the same as each other. Then $\mathbf{B} = \nabla \times \mathbf{A}$ and $\mathbf{E} = -\dot{\mathbf{A}}$, and the classical Hamiltonian from the above for a Maxwellian system becomes

$$H = \frac{1}{2}\int_V d\mathbf{r}\left[\varepsilon \dot{\mathbf{A}}^2(\mathbf{r},t) + \frac{(\nabla \times \mathbf{A}(\mathbf{r},t))^2}{\mu}\right]. \tag{39.4.4}$$

[15] This gauge is valid since it implies that there is no charge in the system since $\nabla \cdot \mathbf{E} = 0$ then.

At this juncture, it is better to define another conjugate variable $\mathbf{\Pi} = \varepsilon \dot{\mathbf{A}}$ so that the above Hamiltonian can be rewritten as

$$H = \frac{1}{2} \int_V d\mathbf{r} \left[\frac{\mathbf{\Pi}^2(\mathbf{r},t)}{\varepsilon} + \frac{(\nabla \times \mathbf{A}(\mathbf{r},t))^2}{\mu} \right]. \tag{39.4.5}$$

The above Hamiltonian is analogous to (39.4.1). In (39.4.1), the conjugate variables p and q vary in tandem to keep the Hamiltonian H a constant. In the above, the conjugate variables $\mathbf{\Pi}$ and $\mathbf{A}$ vary in tandem to a keep H in (39.4.5) a constant. The result is the classical Maxwell's equations [321]. The derivation is rather tedious using functional derivatives and variational calculus. We will skip the derivation and just present the classical Maxwell's equations as given in [336]. They can be derived by Hamiltonian theory to be

$$\nabla \times \mathbf{H}(\mathbf{r},t) = \varepsilon \partial_t \mathbf{E}(\mathbf{r},t) \tag{39.4.6}$$

$$\nabla \times \mathbf{E}(\mathbf{r},t) = -\mu \partial_t \mathbf{H}(\mathbf{r},t) \tag{39.4.7}$$

The above are classical Maxwell's equations for Ampere's law and Faraday's law. The solutions of these equations will maintain the classical Hamiltonian H to be a constant as the electromagnetic fields vary as a function of time.

For simplicity, we look at the 1D case, such that a plane wave is a solution to the classical Maxwell's equations. Hence, we can assume a 1D traveling wave with periodic boundary condition to show that the Hamiltonian above is in fact independent of time. As such, we let $\mathbf{A} = \hat{x} A_x$, and have

$$A_x(z,t) = \frac{1}{2} A_0 e^{-i\omega_l t + i k_l z} + c.c. = A_x^{(+)}(z,t) + A_x^{(-)}(z,t) \tag{39.4.8}$$

where $A_x^{(\pm)}$ represent the positive frequency and negative frequency components of the field, and $k_l = \omega_l / c_0$.[16] They are also the complex conjugate of each other, making the above expression real valued. After the above are squared, and substituted into the Hamiltonian (39.4.4), there will be some oscillatory terms, and constant terms.[17] With periodic boundary condition, the oscillatory terms will integrate to zero, leaving behind only the constant terms. Thus, integrating over the volume such that $\int_V d\mathbf{r} = \mathcal{A} \int_0^L dz$, where L is the exact period, the Hamiltonian (39.4.4) then becomes (see Exercises)

$$H = \frac{V\varepsilon_0}{2} \omega_l^2 \left| \frac{A_l}{2} \right|^2 + \frac{V}{2\mu_0} k_l^2 \left| \frac{A_l}{2} \right|^2 = V \varepsilon_0 \omega_l^2 \left| \frac{A_l}{2} \right|^2. \tag{39.4.9}$$

where $V = \mathcal{A}\mathcal{L}$, is the mode volume.[18] In the above, the two terms above have been combined into one term after noticing that $k_l = \omega_l / c_0$. The above is in fact independent of time, and if the above plane wave is carrying n photons, then the total energy is

$$H = n\hbar\omega_l \tag{39.4.10}$$

[16] For instance, we can choose the period to be L so that $k_l = 2l\pi/L$.

[17] Squaring (39.4.8) will give four terms, two self terms and two cross terms. The constant terms come from the cross terms since their phases cancel, but there are two cross terms.

[18] A sanity check using the $|\mathbf{E}|^2 = \omega^2 |\mathbf{A}|$, the above is equal to $\varepsilon_0 |\mathbf{E}|^2 V$. It includes both the energy stored in the $\mathbf{E}$ and $\mathbf{H}$ fields.

where n is the number of photons in the volume V. By equating this with the above, we deduce that

$$\left|\frac{A_l}{2}\right|^2 = \frac{n\hbar\omega_l}{V\varepsilon_0\omega_l^2} \tag{39.4.11}$$

The above is the semi-classical derivation to relate photon number to the field amplitude.

39.4.2 Hamiltonian—Quantum Picture

We saw in the pendulum case, to get the equation for the quantum pendulum, we elevate the classical conjugate variables, the momentum p and position q in the classical Hamiltonian to become quantum operators. For the electromagnetic oscillation, the classical conjugate variables in the classical Hamiltonian are $\mathbf{\Pi}(\mathbf{r},t)$ and $\mathbf{A}(\mathbf{r},t)$ as shown in (39.4.5). They are analogous to p and q in the pendulum, and they need to be elevated to become quantum operators. To this end, to get the quantum Hamiltonian, we elevate the conjugate variables to become operators in the Heisenberg picture. The quantum Hamiltonian then becomes

$$\hat{H} = \frac{1}{2}\int_V d\mathbf{r}\left[\frac{\hat{\mathbf{\Pi}}^2(\mathbf{r},t)}{\varepsilon} + \frac{1}{\mu}\left(\nabla\times\hat{\mathbf{A}}(\mathbf{r},t)\right)^2\right]. \tag{39.4.12}$$

where $\hat{\mathbf{\Pi}}(\mathbf{r},t) = \varepsilon\dot{\hat{\mathbf{A}}}(\mathbf{r},t)$. In principle, by doing variational calculus, we can derive the quantum Maxwell's equations from the above quantum Hamiltonian, but we will not do that in this course [321]. Instead, we will present the quantum Maxwell's equations below:

$$\nabla\times\hat{\mathbf{H}}(\mathbf{r},t) = \varepsilon\partial_t\hat{\mathbf{E}}(\mathbf{r},t) \tag{39.4.13}$$

$$\nabla\times\hat{\mathbf{E}}(\mathbf{r},t) = -\mu\partial_t\hat{\mathbf{H}}(\mathbf{r},t) \tag{39.4.14}$$

which are quantum Ampere's law and quantum Faraday's law. In the above, $\hat{\mathbf{E}}$ and $\hat{\mathbf{H}}$ are operator representations of the random observables $\mathbf{E}$ and $\mathbf{H}$. For instance, each component of $\mathbf{E}$, namely E_x, E_y, and E_z are now random observables which are represented by operators $\hat{E}_x$, $\hat{E}_y$, and $\hat{E}_z$. The operators themselves are not random, but together with the state vector $|\Psi\rangle$, they endow the relevant observables with random properties.

The quantum Maxwell's equations above are very similar to the classical Maxwell's equation. But by themselves, they are incomplete description of a quantum system because they contain field operators that need to act on a quantum state $|\Psi\rangle$. The quantum state time-evolves according to the quantum state equation

$$\hat{H}|\Psi(t)\rangle = i\hbar\frac{\partial|\Psi(t)\rangle}{\partial t} \tag{39.4.15}$$

[19]

We can easily see that a plane wave multiplied by a position independent operator is a solution to quantum Maxwell's equations in (39.4.13). Thus, focusing on the l-th mode of the solution in

[19] Many authors call this the Schrödinger equation, but it has been applied to all kinds of Hamiltonians including Dirac's Hamiltonian, as well as the electromagnetic Hamiltonian defined in (39.4.12). So we call this the quantum state equation.

free space, we let

$$\hat{A}_x(z,t) = \frac{1}{2}\hat{A}_l e^{-i\omega_l t + ik_l z} + h.c. \tag{39.4.16}$$

where "h.c" stands for "Hermitian conjugate" since we have operators now.[20] We can derive the quantum $\hat{\mathbf{H}}$ and $\hat{\mathbf{E}}$ fields showing that they both satisfy quantum Maxwell's equations above. Furthermore, substituting the above into the quantum Hamiltonian in (39.4.12) where all the fields are elevated to become operators, the quantum Hamiltonian for the l-th mode in (39.4.12) becomes (see Exercises)

$$\hat{H}_l = \frac{1}{4}V\varepsilon_0\omega_l^2(\hat{A}_l\hat{A}_l^\dagger + \hat{A}_l^\dagger\hat{A}_l) \tag{39.4.17}$$

The subscript l underscores that we are looking at the l-th mode Hamiltonian.[21] The above expression ensues because now, $\hat{A}_l$ and $\hat{A}_l^\dagger$ are non-commuting operators. But the above is symmetrized to make the Hamiltonian Hermitian. Next, if we let

$$\frac{1}{2}\sqrt{V\varepsilon_0\omega_l^2}\hat{A}_l = \sqrt{\frac{\hbar\omega_l}{2}}\hat{a}_l \tag{39.4.18}$$

where $\hat{a}_l$ is the annihilation operator for the l-th mode, then the quantum Hamiltonian becomes

$$\hat{H} = \frac{1}{2}\hbar\omega_l\left(\hat{a}_l\hat{a}_l^\dagger + \hat{a}_l^\dagger\hat{a}_l\right) \tag{39.4.19}$$

The above is the quantum Hamiltonian of the l-th mode of the electromagnetic oscillator. Therefore, the above is clearly "homomorphic" to the Hamiltonian of the quantum pendulum as expressed in the previous chapter.

Consequently, the quantum vector potential in (39.4.16) can now be expressed in terms of the annihilation operator $\hat{a}_l$ as

$$\hat{A}_x(z,t) = \sqrt{\frac{\hbar}{2V\varepsilon_0\omega_l}}\hat{a}_l e^{-i\omega_l t + ik_l z} + h.c. \tag{39.4.20}$$

By comparing (39.4.20) and (39.4.16), we see that the annihilation operator $\hat{a}_l(t)$ time evolves as

$$\hat{a}_l(t) = \hat{a}_l(0)e^{-i\omega_l t} \tag{39.4.21}$$

The above is also corroborated by solving the Heisenberg equation of motion for $\hat{a}_l(t)$ as in (38.7.6) [331, eq. (2.21)]. By the same token, we can find the quantum electromagnetic fields for a standing wave between a parallel plate or in a resonant cavity.

39.5 Photon-Carrying Plane Wave

In the quantum world, these oscillators are now converted to quantum oscillators. Instead of carrying a continuum of energy as allowed by classical theory, now these quantum oscillators can

[20] By this symmetrization, the above operator is clearly Hermitian.

[21] The mode Hamiltonians can be considered separately since these modes are orthogonal to each other. The total Hamiltonian of the system is the sum of the Hamiltonian of each individual modes similar to the mode decomposition approach expressed in [337].

convect a finite and discrete amount of energy as espoused by quantum theory. This finite packet of energy is what we call a photon: it is massless, and propagates at the speed of light. It is quantized because each oscillator can only carry quantized energies: it can absorb and release energy in discrete amount.

$$\hat{a}_l e^{-i\omega_l t + ik_l z} \tag{39.5.1}$$

We can regard the above as a "flying" quantum pendulum. In the above, $\hat{a}_l$ is the operator representation of the complex amplitude of the pendulum, while $\hat{n} = \hat{a}_l^\dagger \hat{a}_l$, also the number operator, is the operator representation of the amplitude square of the oscillation. The eigenstates of the $\hat{n}$ are the Fock states $|n\rangle$ with eigenvalue n. From the above, we see that the flying pendulum, $\hat{a}_l e^{-i\omega_l t + ik_l z}$ convects with it a countable finite packet of energy, each of which is the energy of a photon.

In the section, we will let the period in the z direction, L become infinitely large so that the discretizations in k_l and ω_l become infinitesimally small: essentially a Fourier series expansion becomes a Fourier transform. But we will still be looking at one Fourier component for simplicity. To this end, we let $k_l \to k$, $\omega_l \to \omega$, and $\hat{a}_l \to \hat{a}$ subsequently. In other words, the plane wave can exist for a continuum of wave number k and frequency ω and drop their dependency on l.

From (39.4.20), we can write

$$\hat{A}_x(z,t) = \sqrt{\frac{\hbar\omega}{2V\varepsilon_0\omega^2}} e^{i\theta} e^{ikz - i\omega t}\hat{a} + h.c. = \hat{A}_x^{(+)}(z,t) + \hat{A}_x^{(-)}(z,t) \tag{39.5.2}$$

$$\hat{A}_x^{(+)}(z,t) = \sqrt{\frac{\hbar\omega}{2V\varepsilon_0\omega^2}} e^{i\theta} e^{ikz - i\omega t}\hat{a} \tag{39.5.3}$$

$$\hat{A}_x^{(-)}(z,t) = \sqrt{\frac{\hbar\omega}{2V\varepsilon_0\omega^2}} e^{-i\theta} e^{-ikz + i\omega t}\hat{a}^\dagger \tag{39.5.4}$$

where $\hat{A}_x^{(+)}(z,t)$ and $\hat{A}_x^{(-)}(z,t)$ are the positive frequency and negative part of the signal $\hat{A}_x(z,t)$. They are also the Hermitian conjugate of each other to ensure that $\hat{A}_x(z,t)$ is Hermitian with real expectation value so that it is the operator representation of a real observable. In this manner, we can keep track of the number of photons riding on the plane wave in the quantum system in the quantum world. In the limit of a large number of photons, we should retrieve the classical field.

In addition, we have four sanity checks to ensure if we are on the right track:

- The proposed field operator has to satisfy quantum Maxwell's equations .
- The quantum observable (represented by quantum operator) together with the quantum state implies $\hat{A}_x(z,t)$ has a random variable $A_x(z,t)$ associated with it. But the quantum observable $\hat{A}_x(z,t)$ by itself does not have random properties. As shall be shown, it is the quantum state that endows the observable with random properties.
- The quantum state time-evolves according to the quantum state equation (39.4.15) where the Hamiltonian is defined by (39.4.12).
- The operators representing these quantum observables can be non commuting. This is their departure from classical observables which are scalar variables, and hence, are commuting variables.

It is common to denote a photon number state called Fock state by $|n\rangle$ which is the abbreviated way to denote Ψ_n that we have introduced early. Using this Fock state, one finds the expectation value of the operator $\hat{A}_x^{(-)}(z,t)\hat{A}_x^{(+)}(z,t)$. To do so, we need to evaluate

$$\langle n|\hat{A}_x^{(-)}(z,t)\hat{A}_x^{(+)}(z,t)|n\rangle \tag{39.5.5}$$

Upon using the above expressions for $\hat{A}_x^{(\pm)}$, we have

$$\langle n|\hat{A}_x^{(-)}(z,t)\hat{A}_x^{(+)}(z,t)|n\rangle = \frac{\hbar\omega}{2V\varepsilon_0\omega^2}\langle n|\hat{a}^\dagger\hat{a}|n\rangle = \frac{n\hbar\omega}{2V\varepsilon_0\omega^2} \tag{39.5.6}$$

In the above, we have made use of the fact that $\hat{a}^\dagger\hat{a} = \hat{n}$, the number operator. Also, the Fock state $|n\rangle$ is an eigenstate of the number operator $\hat{n}$ such that $\hat{n}|n\rangle = \hat{a}^\dagger\hat{a}|n\rangle = n|n\rangle$ with eigenvalue n. As can be seen, the correspondence principle is satisfied: in the limit of large number of photons, the quantum case agrees with the classical case.There is a factor of 2 difference because $\hat{A}_x^{(-)}(z,t)\hat{A}_x^{(+)}(z,t)$ corresponds to only the energy stored in the electric field which is only half the Hamiltonian.

39.6 Wave of Arbitrary Polarization—Quantum Case

In general, when both horizontal and vertical polarizations of the plane wave are present, we can write (ignoring the arbitrary phase)

$$\hat{\mathbf{A}}(\mathbf{r},t) = \sum_{s\in\{v,h\}}\sqrt{\frac{\hbar}{2\omega\varepsilon_0 V}}\mathbf{e}_s\hat{a}_s e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + h.c. \tag{39.6.1}$$

where $\mathbf{e}_s$ is $\mathbf{e}_v$ for vertical polarization and is $\mathbf{e}_h$ for horizontal polarization. Therefore, there are actually two quantum oscillators working in unison in the above equation: one oscillating horizontally, and the other one oscillating vertically. With this kind of oscillators, we will illustrate how it can be used for quantum communication.

Alternatively, we rewrite the above as

$$\hat{\mathbf{A}}(\mathbf{r},t) = \sum_{s\in\{v,h\}}\sqrt{\frac{\hbar}{2\omega\varepsilon_0 V}}\mathbf{e}_s\hat{a}_s e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + h.c. = \hat{\mathbf{A}}^{(+)}(\mathbf{r},t) + \hat{\mathbf{A}}^{(-)}(\mathbf{r},t) \tag{39.6.2}$$

where

$$\hat{\mathbf{A}}^{(+)}(\mathbf{r},t) = \sum_{s\in\{v,h\}}\sqrt{\frac{\hbar}{2\omega\varepsilon_0 V}}\mathbf{e}_s\hat{a}_s e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{39.6.3}$$

$$\hat{\mathbf{A}}^{(-)}(\mathbf{r},t) = \sum_{s\in\{v,h\}}\sqrt{\frac{\hbar}{2\omega\varepsilon_0 V}}\mathbf{e}_s\hat{a}_s^\dagger e^{-i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{39.6.4}$$

where $\hat{\mathbf{A}}^{(-)}(\mathbf{r},t) = \left[\hat{\mathbf{A}}^{(+)}(\mathbf{r},t)\right]^\dagger$.

To show the randomness of the field, we assume that the two polarizations above are in perfect phase coherence but oscillating in two different quantum oscillators. In fact, there are two Hamiltonians in the system, each for the orthogonal polarizations of the plane wave, viz.,

$$\hat{H} = \hat{H}_v + \hat{H}_h \tag{39.6.5}$$

In the above, $\hat{H}_v$ and $\hat{H}_h$ are the Hamiltonians for the v-pol oscillation and h-pol oscillations respectively. Therefore, there are two quantum oscillators in the system. For simplicity, we assume that there is only one photon present in the system even though there are two oscillators. Hence, this photon, in this case, the photon is in a quantum linear superposition two states or polarizations. Or one can think that the photon can "morph" seamlessly between the two polarizations. The measurement outcome of the photon polarization is randomly found in one of these two polarizations[22] To facilitate this, the two oscillators have to be in perfect phase coherence. Since there are two Hamiltonians for the oscillators, each Hamiltonian has its own number state or Fock state. Let us look at the one-photon Fock state more carefully. The products of one photon Fock states are:[23]

$$|1_v\rangle|0_h\rangle, \quad |0_v\rangle|1_h\rangle \Rightarrow |\Psi_{\text{one-photon}}\rangle = A_v|1_v\rangle|0_h\rangle + A_h|0_v\rangle|1_h\rangle = A_v|1_v\rangle + A_h|1_h\rangle \tag{39.6.6}$$

The last symbols are the most reticent ones where the zero photon states implied and are dropped. When we have two quantum systems working in unison or in coherence, the basis state vectors for the quantum states are the products of the individual basis vectors.

Since the above has probabilistic interpretation, then $|A_v|^2 + |A_h|^2 = 1$. What is measurable is the expectation of a quantum operator with respect to the quantum state, which is $|\Psi_{\text{one-photon}}\rangle$ in this case.

However, it is well-known that the expectation value of the field operator $\hat{\mathbf{A}}(\mathbf{r}, t)$ is zero with respect to the Fock states or the photon number states. Hence, the Fock states are termed nonclassical states. However, the expectation value of the energy, which is proportional to $\hat{\mathbf{A}}^{(-)} \cdot \hat{\mathbf{A}}^{(+)}$, is nonzero, because it is in turn proportional to $\hat{a}^\dagger \hat{a}$ which is the number operator. Thus, the expectation value of the above operator with respect to the one-photon state is nonzero. This is given by

$$M = \langle \Psi_{\text{one-photon}} | \hat{\mathbf{A}}^{(-)}(\mathbf{r}, t) \cdot \hat{\mathbf{A}}^{(+)}(\mathbf{r}, t) | \Psi_{\text{one-photon}} \rangle \tag{39.6.7}$$

The above is something measurable in the laboratory. But this is a random variable because the photon chooses to ride randomly between the two polarizations.[24] The above gives the expectation value or the average of this random variable.

For clarity, after simplification using (39.6.3) and (39.6.4), we define the operator explicitly as

$$\hat{O} = \hat{\mathbf{A}}^{(-)}(\mathbf{r}, t) \cdot \hat{\mathbf{A}}^{(+)}(\mathbf{r}, t) = \frac{\hbar}{2\omega\varepsilon_0 V} \left[\hat{a}_h^\dagger \hat{a}_h + \hat{a}_v^\dagger \hat{a}_v \right] \tag{39.6.8}$$

In arriving at the above, we have made use of that $\mathbf{e}_v$ and $\mathbf{e}_h$ are orthogonal to each other in the dot product above. Continuing to work on the algebra explicitly, we have

$$M = |A_v|^2 C \langle 1_v | \hat{a}_v^\dagger \hat{a}_v | 1_v \rangle + |A_h|^2 C \langle 1_h | \hat{a}_h^\dagger \hat{a}_h | 1_h \rangle = \left(|A_v|^2 + |A_h|^2 \right) C = C \tag{39.6.9}$$

[22] The random nature of the outcome is more to do with the quantum interpretation of the measurement.

[23] These are outer products or tensor products sometimes denoted as $|1_v\rangle \otimes |0_h\rangle$, but physicists love the more succinct version used here [338, eq. (21.3)] [78, sec. 12.6.3].

[24] Only ghost and angel can do that. We may call the random state the 'ghost-angel' state.

where $C = \frac{\hbar}{2\omega\varepsilon_0 V}$ which is commensurate with (39.5.6). In the above, we have made use of that $|A_v|^2 + |A_h|^2 = 1$ and that $\hat{a}_s^\dagger \hat{a}_s |1_s\rangle = |1_s\rangle$ where $s \in (v, h)$.

The physical interpretation of the above is that the system launch a quantum state $|\Psi_{\text{one-photon}}\rangle$ which is a random superposition of two one-photon states, $|1_v\rangle$ and $|1_h\rangle$. This one-photon state is being operated upon by the $\hat{O}$ operator, and the outcome is then projected onto the one-photon state $\langle\Psi_{\text{one-photon}}|$. In quantum theory, the projection embeds the concept of quantum measurement where the quantum state collapses into the measured state.

From the mathematics, we see two possible paths of quantum collapse. If the measurement outcome is the state $|1_h\rangle$, then the probability of this outcome is proportional to

$$\langle 1_h|\hat{\mathbf{A}}^{(-)}(\mathbf{r},t)\cdot\hat{\mathbf{A}}^{(+)}(\mathbf{r},t)|\Psi_{\text{one-photon}}\rangle = \frac{\hbar\omega}{2V\varepsilon_0\omega^2}|A_h|^2 \tag{39.6.10}$$

If the measurement outcome is the state with $|1_v\rangle$, its value is proportional to

$$\langle 1_v|\hat{\mathbf{A}}^{(-)}(\mathbf{r},t)\cdot\hat{\mathbf{A}}^{(+)}(\mathbf{r},t)|\Psi_{\text{one-photon}}\rangle = \frac{\hbar\omega}{2V\varepsilon_0\omega^2}|A_v|^2 \tag{39.6.11}$$

Since $|A_h|^2 + |A_v|^2 = 1$, the probability of detecting a photon is $|A_h|^2$ and $|A_v|^2$ respectively for a horizontally and vertically polarized photon. The above is completely random!

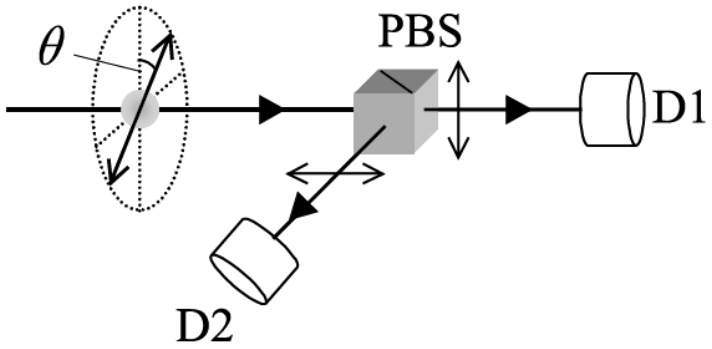


Figure 39.4: Polarization of the incident photon can be detected with a polarizing beam splitter (PBS) and single photon detectors (from Quantum Optics, M. Fox [332]). In quantum theory, the polarization is random.

The fact that a photon state remains *incognito* before a measurement can be used for quantum encryption or quantum communication. There is even a non-cloning theorem that guarantees that quantum communication is secure. If Alice and Bob[25] were to exploit the random properties of quantum theory to communicate information buried in a photon, Eve, an eave-dropper, cannot steal a photon, alter its property, and re-send it along and yet remains undiscovered [77]!

[25]Two proverbial stars in quantum communications [77].

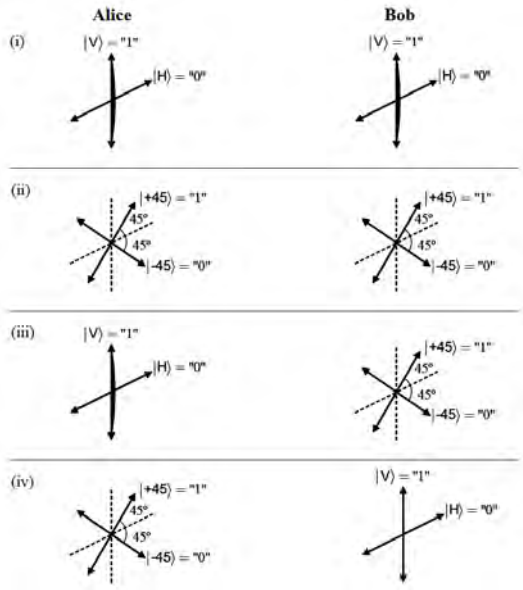


Figure 39.5: Communication between Alice and Bob using single-photon source and simplified polarizer measurement schemes (from DAB Miller [77]).

39.6.1 Polychromatic Photons vs Monochromatic Photons

In the above, we have assume that the photon is associated with a plane wave that is entirely monochromatic (CW or time-harmonic in other parlances). But a CW signal is unphysical as it has no beginning nor end. Photons in the laboratory have been associated with a flight time, and this is well defined only if the electromagnetic pulse associated with a photon is a localized pulse in time and space. We can think of a photon "riding" on a wave, and this is only possible and agrees with experimental observation if the photon rides on a wave of finite duration, or a localized pulse. Therefore, the electromagnetic pulse that a photon rides on has to be localized.

We can also imagine how a photon is emitted from an atom: The atom experiences an electronic transition of its energy level from an upper level to a lower level. Due to energy conservation, a photon has to be emitted that is equal to the energy gap of the transition. A photon can be thought of as a pulse of electromagnetic energy that carries energy from the atom to infinity for energy conservation purpose. This pulse of energy has a beginning and an end: hence, it is broadband and causal. Only a photon with polychromatic field can carry this photon away from the atom. A polychromatic field can be expressed as

$$\hat{A}_x(z,t) = \sum_l \sqrt{\frac{\hbar}{2V\varepsilon_0\omega_l}}\hat{a}_l e^{-i\omega_l t + ik_l z} + h.c. \tag{39.6.12}$$

where l is used to index the frequency of the field or mode. The above can be thought of as a multi-modal field where each propagating field is a mode.

The corresponding Hamiltonian for the multi-modal field is then

$$\hat{H} = \sum_{l=1}^{N} \hat{H}_l \tag{39.6.13}$$

39.6.2 Quantum States of a Multimodal Field

In order to solve the quantum state equation, we need to find the space that (39.6.13) would operate on. The field lives in a Hilbert space that is the outer product of individual field modes of the system. Assume that the i-th mode lives in the Hilbert space $\mathscr{E}_i$, then the above Hilbert space for (39.6.13) is the outer product of these spaces, viz., [339, p. 175]

$$\mathscr{E}_r = \mathscr{E}_1 \otimes \mathscr{E}_2 \cdots \otimes \mathscr{E}_i \otimes \cdots \otimes \mathscr{E}_N \tag{39.6.14}$$

We can span each of the spaces with their Fock states. For example, $\{|n_i\rangle, n_i = 0, \ldots, \infty\}$ spans the space $\mathscr{E}_i$. For simplicity, we will denote such a spanning basis set as $\{|n_i\rangle\}$ with n_i running from 0 to ∞ implied.

Then the set of spanning basis for the above Hilbert space consisting of N mode is

$$\{|n_1\rangle, |n_2\rangle, \ldots, |n_i\rangle, \ldots, |n_N\rangle\} = \{|n_1, n_2, \ldots, n_i, \ldots, n_N\rangle\} \tag{39.6.15}$$

spans the multi-modal field space. (The second form above is the abbreviated form.) A linear superposition of the above can be used to approximate a function (or a vector, or state vector) in $\mathscr{E}_r$. In principle, N can be infinitely large.

Humongous Hilbert Space

The above Hilbert space is infinitely large. For practical computations, we would truncate the size of the Hilbert space with approximations. If we have a cavity with N modes, and if each mode is spanned with the spanning basis for M Fock states, the number of spanning basis is proportional to M^N (see [340] and references therein). This is provided that all the modes are phase coherent, and that a photon can morph from one mode to the other modes seamlessly. Whether all these modes can be maintained to be phase coherent has to be verified by experiments!

Figure 39.6: The ability of a photon to move seamlessly from modes to modes in a resonant cavity reminds me of a monkey being able to easily jump from trees to trees in a tropical jungle! It shows the beauty of Nature (courtesy of Wikipedia).

.

Abbreviated Notations

The above notation is cumbersome. Abbreviated notations are often used. One such abbreviation is that [341, 342]

$$|\{n_k\}\rangle = |n_{k_1}\rangle|n_{k_2}\rangle|n_{k_2}\rangle \ldots = \Pi_k|n_k\rangle \tag{39.6.16}$$

For a single-photon state occupying mode k, it should rightfully be denoted as

$$|0, \ldots, 1_k, \ldots, 0\rangle \tag{39.6.17}$$

As an abbreviation, it can be denoted as [338,] [78,]

$$|1\rangle_k \tag{39.6.18}$$

A single-photon state can also be written as a linear superposition of many sigle-photon states as follows:

$$|\Phi^{(1)}\rangle = \sum_{p=1}^{N} \tilde{w}_p|1\rangle_p = \sum_{p=1}^{N} \tilde{w}_p \hat{a}_p^\dagger|0\rangle = \hat{\mathbf{a}}^\dagger \cdot \tilde{\mathbf{w}}|0\rangle \tag{39.6.19}$$

where $[\tilde{\mathbf{w}}]_p = \tilde{w}_p$ is the probability amplitude of $|1\rangle_p$ that incorporates the spectral amplitude of the wave packet, with the normalization condition $\tilde{\mathbf{w}}^\dagger \cdot \tilde{\mathbf{w}} = 1$. The above implies that a single photon is "riding" on the localized wave packet. The weights $\tilde{w}_p$ can be Gaussian tapered so that we get a Gaussian pulse in mode space as well as the coordinate space. These pulses have been used successfully to study the Hong-Ou-Mandel effect [343].

39.7 Epilogue

In conclusion, the quantum theory of light is a rather complex subject. It cannot be taught in just two lectures, but what we wish is to give you a taste of this theory. One needs to read and re-read these two chapters to imbibe these concepts. It takes much longer to learn this subject well: after all, it is the by product of almost a century of intellectual exercise. This knowledge is still very much in its infancy. Hopefully, the more we teach this subject, the better we can articulate, understand, and explain the ideas behind this subject. When James Clerk Maxwell completed the theory of electromagnetics over 150 years ago, and wrote a book on the topic, rumor has it that most people could not read beyond the first 50 pages of his tome [18]. But after over a century and a half of regurgitation, we can now teach the subject to undergraduate students! When Maxwell put his final stroke to the equations named after him, he could never have foreseen that these equations are valid from sub-atomic lengthscales to galactic lengthscales, and from static to ultra-violet frequencies. Now, these equations are even valid from classical to the quantum world as well!

Hopefully, by introducing these frontier knowledge in electromagnetic field theory in this course, it will pique your interest enough in this subject, so that you will take this as a life-long learning experience. I suggest that you read these two chapters over and again so that this knowledge can sink into your soul!

If technology is the gift of God, may God give us the inspiration to use it wisely!

Exercises for Lecture 39

Problem 39-1: This problem refers to Chapter 39.

(i) Show that C_0 in (39.2.11) is as indicated.

(ii) Derive (39.3.15) of the text.

(iii) Derive (39.3.17).

(iv) Derive (39.4.9) of the text.

(v) Derive (39.4.17).

(vi) Derive (39.6.10) and (39.6.11).

(vii) Explain why the humongous Hilbert space scales as M^N.

Appendix A

A.1 Meaning of the Function of an Operator

We will digress here to discuss the meaning of a function of an operator, which occurs in (39.2.14): The exponential function is a function of the operator $\hat{H}$. This knowledge is used prevalently in control theory, but not so much in other parts of electrical engineering. This is best explained by expanding the pertinent function into a Taylor series, namely, a polynomial series given as

$$f(\overline{\mathbf{A}}) = f(0)\overline{\mathbf{I}} + f'(0)\overline{\mathbf{A}} + \frac{1}{2!}f''(0)\overline{\mathbf{A}}^2 + \ldots + \frac{1}{n!}f^{(n)}(0)\overline{\mathbf{A}}^n + \ldots \tag{A.1.1}$$

Without loss of generality, we have used the matrix operator $\overline{\mathbf{A}}$ as an illustration. The above series has no meaning unless it acts on an eigenvector of the matrix operator $\overline{\mathbf{A}}$, where $\overline{\mathbf{A}}\mathbf{v} = \lambda\mathbf{v}$. Hence, by applying the above equation (A.1.1) to an eigenvector $\mathbf{v}$ of $\overline{\mathbf{A}}$, we have

$$\begin{aligned} f(\overline{\mathbf{A}})\mathbf{v} &= f(0)\mathbf{v} + f'(0)\overline{\mathbf{A}}\mathbf{v} + \frac{1}{2!}f''(0)\overline{\mathbf{A}}^2\mathbf{v} + \ldots + \frac{1}{n!}f^{(n)}(0)\overline{\mathbf{A}}^n\mathbf{v} + \ldots \\ &= f(0)\mathbf{v} + f'(0)\lambda\mathbf{v} + \frac{1}{2!}f''(0)\lambda^2\mathbf{v} + \ldots + \frac{1}{n!}f^{(n)}(0)\lambda^n\mathbf{v} + \ldots = f(\lambda)\mathbf{v} \end{aligned} \tag{A.1.2}$$

The last equality follows by re-summing the Taylor series back into a function. Applying this to an exponential function of an operator, we have, when $\mathbf{v}$ is an eigenvector of $\overline{\mathbf{A}}$, that

$$e^{\overline{\mathbf{A}}}\mathbf{v} = e^{\lambda}\mathbf{v} \tag{A.1.3}$$

The above proof applies to any polynomial function $f(x)$. In a word,

$$f(\overline{\mathbf{A}})\mathbf{v} = f(\lambda)\mathbf{v} \tag{A.1.4}$$

For a general vector $\mathbf{V}$, it can be expanded as a sum of the eigenvectors of $\overline{\mathbf{A}}$, namely, that

$$\mathbf{V} = \sum_i a_i\mathbf{v}_i \tag{A.1.5}$$

where $\mathbf{v}_i$ is the i-th eigenvector of $\overline{\mathbf{A}}$. Then one can proceed to evaluate in the general case of

$$f(\overline{\mathbf{A}})\mathbf{V} = \sum_i a_i f(\overline{\mathbf{A}})\mathbf{v}_i = \sum_i a_i f(\lambda_i)\mathbf{v}_i \tag{A.1.6}$$

for arbitrary $\mathbf{V}$. Here, λ_i is the i-th eigenvalue of $\overline{\mathbf{A}}$.

A.2 Resolution of Identity Operator

It is prudent to introduce the beautiful operator called the resolution of identity operator in here [317]. Such an operator is useful in a number of manipulations of quantum equations involving operators.

Matrix Elements and Matrix Representations

Before proceeding, we would like to introduce the concepts of matrix elements and representations. Given an operator $\mathscr{L}$, its matrix element, in Dirac notation, is given by

$$\left[\overline{\mathbf{L}}\right]_{mn} = L_{mn} = \langle \Psi_m | \mathscr{L} | \Psi_n \rangle \tag{A.2.1}$$

The matrix $\overline{\mathbf{L}}$ is the matrix representation of the operator $\mathscr{L}$. This is similar to that of computational electromagnetics in Section 36.6.3 except that we have use Dirac notation here. Also, in quantum world, the operators are invariably Hermitian.

Assume that an N dimensional linear vector space is spanned by N orthonormal unit vectors

$$\{\mathbf{e}_i, i = 1, \ldots, N\} \tag{A.2.2}$$

Then it is quite clear that in an N dimensional linear vector space, the identity matrix can be expanded in terms of the outer products of these unit vectors as

$$\bar{\mathbf{I}} = \sum_{i=1}^{N} \mathbf{e}_i \otimes \mathbf{e}_i = \sum_{i=1}^{N} \mathbf{e}_i \mathbf{e}_i^{\dagger}. \tag{A.2.3}$$

The above can be generalized to an infinite dimensional Hilbert space using the orthonormal eigenvectors of a Hermitian operator. For instance, for the operator $\hat{q}$ such that

$$\hat{q}|q\rangle = q|q\rangle \tag{A.2.4}$$

where $|q\rangle$ is an orthonormalized eigenvector with eigenvalue q, we can use $|q\rangle$ as the unit vector for coordinate space for spanning the Hilbert space. As an analogous to (A.2.3), we can define an identity operator

$$\hat{I} = \int_{-\infty}^{\infty} dq |q\rangle\langle q|, \tag{A.2.5}$$

Similar comment applies to the $\hat{p}$ operator. Namely,

$$\hat{p}|p\rangle = p|p\rangle \tag{A.2.6}$$

Then an identity operator, using $|p\rangle$ orthonormal basis vectors, is

$$\hat{I} = \int_{-\infty}^{\infty} dp |p\rangle\langle p|, \tag{A.2.7}$$

Switching Representations in Different Basis in Coordinate and Momentum Domains

The identity operator can be inserted in an inner product without changing its value. Therefore, using it, we can switch the representations of the operators in different domains easily.

Using the resolution of the identity operator, we can rewrite

$$\langle q\rangle = \langle\Psi|\hat{q}|\Psi\rangle = \int\int dqdq'\langle\Psi|q'\rangle\langle q'|\hat{q}|q\rangle\langle q|\Psi\rangle = \int\int dqdq'\Psi^*(q')\langle q'|\hat{q}|q\rangle\Psi(q) \tag{A.2.8}$$

where we have defined $\langle q|\Psi\rangle$ as the function $\Psi(q)$ etc. Since $|q\rangle$ is an eigenvector of $\hat{q}$, with eigenvalue q, then $\hat{q}|q\rangle = q|q\rangle$ and

$$\langle q'|\hat{q}|q\rangle = q\langle q'|q\rangle = \delta(q'-q)q \tag{A.2.9}$$

the above then becomes

$$\langle q\rangle = \int dq\Psi^*(q)q\Psi(q) \tag{A.2.10}$$

The above is the same as what we had before in (38.4.3) proving the correctness of the notation there. Using the resolution of identity operator,

In the above, $\hat{q}$ is an abstract operator, with eigenvector $|q\rangle$ that spans the infinite dimensional Hilbert space. But $\langle q'|\hat{q}|q\rangle = \delta(q'-q)q$ is the matrix representation of the operator $\hat{q}$ in the space spanned by the basis vector $|q\rangle$.

We can repeat the above exercise for finding $\langle p\rangle$ in the coordinate domain.

$$\langle p\rangle = \langle\Psi|\hat{p}|\Psi\rangle = \int\int dqdq'\langle\Psi|q'\rangle\langle q'|\hat{p}|q\rangle\langle q|\Psi\rangle = \int\int dqdq'\Psi^*(q')\langle q'|\hat{p}|q\rangle\Psi(q) \tag{A.2.11}$$

We could also have calculated the above using the definition of $\hat{p}$ in coordinate domain, which is usually given sloppily as

$$\hat{p} = -i\hbar\frac{d}{dq} \tag{A.2.12}$$

From (A.2.11), we deduce that, similar to Section 38.4.1, we have that

$$\langle p\rangle = \langle\Psi|\hat{p}|\Psi\rangle = \int dq\Psi^*(q)\left(-i\hbar\frac{d}{dq}\right)\Psi(q) \tag{A.2.13}$$

Using that, and comparing the two equations, (A.2.11) and (A.2.13), we deduce that

$$\langle q'|\hat{p}|q\rangle = -i\hbar\delta(q'-q)\frac{d}{dq} \tag{A.2.14}$$

It is often said that the operator $\hat{p} = -i\hbar\frac{d}{dq}$. Strictly speaking, this is not koshered. What is implied is that the diagonal elements of the matrix representation in coordinate basis of the operator $\hat{p}$ is $-i\hbar\frac{d}{dq}$. A matrix representation (or element) of an operator always has two indices, q' and q in this case as in (A.2.14).

A.3 Density Matrix or Operator

An elegant way to represent a quantum state is via the density operator or density matrix. It is defined as

$$\hat{\rho} = |\Psi\rangle\langle\Psi| \tag{A.3.1}$$

It has the same information as the state vector $|\Psi\rangle$, but only relative phases are important, as $\langle\Psi|$ is the conjugate transpose of $|\Psi\rangle$. We shall show that the expectation of the operator $\hat{A}$ in this state $|\Psi\rangle$ is

$$\langle\hat{A}\rangle = \langle\Psi|\hat{A}|\Psi\rangle = \text{tr}\left(\hat{A}\hat{\rho}\right) \tag{A.3.2}$$

Therefore, knowing the density operator $\hat{\rho}$ is equivalent to knowing the quantum state $|\Psi\rangle$ of a system.

We can calculate the trace of an operator in terms of the sum of the diagonal elements of its matrix representation. Assume that we have a countable set of orthonormal basis vectors that is complete, we can find the matrix representation of the operator $\hat{A}\hat{\rho}$ so that its trace can be found. The matrix elements or representation of the operator $\hat{A}\hat{\rho}$ is given by

$$M_{mn} = \langle\phi_m|\hat{A}\hat{\rho}|\phi_n\rangle \tag{A.3.3}$$

Hence, the trace of the matrix can be found as

$$\text{tr}\left(\hat{A}\hat{\rho}\right) = \sum_n \langle\phi_n|\hat{A}\hat{\rho}|\phi_n\rangle \tag{A.3.4}$$

We show that the above is basis independent by inserting the identity operator $\hat{I} = \sum_m |\phi_m\rangle\langle\phi_m|$ twice in the inner product on the right-hand side of (A.3.2), to yield[1]

$$\langle\Psi|\hat{A}|\Psi\rangle = \sum_{n,m} \langle\Psi|\phi_n\rangle\langle\phi_n|\hat{A}|\phi_m\rangle\langle\phi_m|\Psi\rangle \tag{A.3.5}$$

Since the factors in the summand are scalars, we can reorder them to give

$$\langle\Psi|\hat{A}|\Psi\rangle = \sum_{n,m} \langle\phi_m|\Psi\rangle\langle\Psi|\phi_n\rangle\langle\phi_n|\hat{A}|\phi_m\rangle = \sum_{n,m} \rho_{mn}A_{nm} = \text{tr}\left(\overline{\boldsymbol{\rho}}\cdot\overline{\mathbf{A}}\right) \tag{A.3.6}$$

where ρ_{mn} is the matrix representation of $\hat{\rho}$ while A_{nm} is the matrix representation of $\hat{A}$. The above is clearly independent of the orthonormal basis we have chosen. Also, it is quite easy to show that $\text{tr}\left(\overline{\boldsymbol{\rho}}\cdot\overline{\mathbf{A}}\right) = \text{tr}\left(\overline{\mathbf{A}}\cdot\overline{\boldsymbol{\rho}}\right)$. Hence, we can write the expectation value as (A.3.2) which is also valid for a continuum basis vectors.

[1] Again, we assume that the set of vectors $|\phi_m\rangle$ forms a complete set.

Bibliography

[1] J. A. Kong, *Theory of electromagnetic waves*. New York, Wiley-Interscience, 1975.

[2] A. Einstein *et al.*, "On the electrodynamics of moving bodies," *Annalen der Physik*, vol. 17, no. 891, p. 50, 1905.

[3] P. A. M. Dirac, "The quantum theory of the emission and absorption of radiation," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 114, no. 767, pp. 243–265, 1927.

[4] R. J. Glauber, "Coherent and incoherent states of the radiation field," *Physical Review*, vol. 131, no. 6, p. 2766, 1963.

[5] C.-N. Yang and R. L. Mills, "Conservation of isotopic spin and isotopic gauge invariance," *Physical review*, vol. 96, no. 1, p. 191, 1954.

[6] G. t'Hooft, *50 years of Yang-Mills theory*. World Scientific, 2005.

[7] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*. Princeton University Press, 2017.

[8] F. Teixeira and W. C. Chew, "Differential forms, metrics, and the reflectionless absorption of electromagnetic waves," *Journal of Electromagnetic Waves and Applications*, vol. 13, no. 5, pp. 665–686, 1999.

[9] W. C. Chew, E. Michielssen, J.-M. Jin, and J. Song, *Fast and efficient algorithms in computational electromagnetics*. Artech House, Inc., 2001.

[10] A. Volta, "On the electricity excited by the mere contact of conducting substances of different kinds. in a letter from Mr. Alexander Volta, FRS Professor of Natural Philosophy in the University of Pavia, to the Rt. Hon. Sir Joseph Banks, Bart. KBPR S," *Philosophical transactions of the Royal Society of London*, no. 90, pp. 403–431, 1800.

[11] A.-M. Ampère, *Exposé méthodique des phénomènes électro-dynamiques, et des lois de ces phénomènes*. Bachelier, 1823.

[12] ——, *Mémoire sur la théorie mathématique des phénomènes électro-dynamiques uniquement déduite de l'expérience: dans lequel se trouvent réunis les Mémoires que M. Ampère a communiqués à l'Académie royale des Sciences, dans les séances des 4 et 26 décembre 1820, 10 juin 1822, 22 décembre 1823, 12 septembre et 21 novembre 1825*. Bachelier, 1825.

[13] B. Jones and M. Faraday, *The life and letters of Faraday*. Cambridge University Press, 2010, vol. 2.

[14] G. Kirchhoff, "Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird," *Annalen der Physik*, vol. 148, no. 12, pp. 497–508, 1847.

[15] L. Weinberg, "Kirchhoff's' third and fourth laws'," *IRE Transactions on Circuit Theory*, vol. 5, no. 1, pp. 8–30, 1958.

[16] T. Standage, *The Victorian Internet: The remarkable story of the telegraph and the nineteenth century's online pioneers*. Phoenix, 1998.

[17] J. C. Maxwell, "A dynamical theory of the electromagnetic field," *Philosophical transactions of the Royal Society of London*, no. 155, pp. 459–512, 1865.

[18] P. J. Nahin, *Oliver Heaviside: sage in solitude*. IEEE Press New York, 1987.

[19] H. Hertz, "On the finite velocity of propagation of electromagnetic actions," *Electric Waves*, vol. 110, 1888.

[20] M. Romer and I. B. Cohen, "Roemer and the first determination of the velocity of light (1676)," *Isis*, vol. 31, no. 2, pp. 327–379, 1940.

[21] Wikipedia, "Newton's Rings," https://en.wikipedia.org/wiki/Newton's_rings.

[22] A. Arons and M. Peppard, "Einstein's proposal of the photon concept–a translation of the Annalen der Physik paper of 1905," *American Journal of Physics*, vol. 33, no. 5, pp. 367–374, 1965.

[23] A. Pais, "Einstein and the quantum theory," *Reviews of Modern Physics*, vol. 51, no. 4, p. 863, 1979.

[24] M. Planck, "On the law of distribution of energy in the normal spectrum," *Annalen der physik*, vol. 4, no. 553, p. 1, 1901.

[25] A. Houck, D. Schuster, J. Gambetta, J. Schreier, B. Johnson, J. Chow, L. Frunzio, J. Majer, M. Devoret, S. Girvin *et al.*, "Generating single microwave photons in a circuit," *Nature*, vol. 449, no. 7160, pp. 328–331, 2007.

[26] Z. Peng, S. De Graaf, J. Tsai, and O. Astafiev, "Tuneable on-demand single-photon source in the microwave range," *Nature communications*, vol. 7, p. 12588, 2016.

[27] B. D. Gates, Q. Xu, M. Stewart, D. Ryan, C. G. Willson, and G. M. Whitesides, "New approaches to nanofabrication: molding, printing, and other techniques," *Chemical reviews*, vol. 105, no. 4, pp. 1171–1196, 2005.

[28] D. J. Griffiths and D. F. Schroeter, *Introduction to quantum mechanics*. Cambridge University Press, 2018.

[29] J. S. Bell, "The debate on the significance of his contributions to the foundations of quantum mechanics, Bell's Theorem and the Foundations of Modern Physics (A. van der Merwe, F. Selleri, and G. Tarozzi, eds.)," 1992.

[30] C. Pickover, *Archimedes to Hawking: Laws of science and the great minds behind them*. Oxford University Press, 2008.

[31] R. Resnick, J. Walker, and D. Halliday, *Fundamentals of physics*. John Wiley, 1988.

[32] J. L. De Lagrange, "Recherches d'arithmétique," *Nouveaux Mémoires de l'Académie de Berlin*, 1773.

[33] S. Ramo, J. R. Whinnery, and T. Duzer van, *Fields and waves in communication electronics, Third Edition*. John Wiley & Sons, Inc., 1995, also 1965, 1984.

[34] J. A. Kong, *Electromagnetic Wave Theory*. EMW Publishing, 2008, also 1985.

[35] H. M. Schey, *Div, grad, curl, and all that: an informal text on vector calculus*. WW Norton New York, 2005.

[36] M. N. Sadiku, *Elements of electromagnetics*. Oxford University Press, 2014.

[37] W. C. Chew, *Waves and fields in inhomogeneous media*. IEEE Press, 1995, also 1990.

[38] A. R. Choudhuri, *The physics of fluids and plasmas: an introduction for astrophysicists*. Cambridge University Press, 1998.

[39] V. J. Katz, "The history of Stokes' theorem," *Mathematics Magazine*, vol. 52, no. 3, pp. 146–156, 1979.

[40] J. C. Maxwell, *A Treatise on Electricity and magnetism*. Dover New York, 1954, first published in 1873, vol. 1 and 2.

[41] A. D. Yaghjian, "Reflections on Maxwell's treatise," *Progress In Electromagnetics Research*, vol. 149, pp. 217–249, 2014.

[42] B. J. Hunt, *The Maxwellians*. Cornell University Press, 2005.

[43] J. C. Maxwell, "Poems of James Clerk Maxwell," https://mypoeticside.com/poets/james-clerk-maxwell-poems.

[44] W. K. Panofsky and M. Phillips, *Classical electricity and magnetism*. Courier Corporation, 2005.

[45] T. Lancaster and S. J. Blundell, *Quantum field theory for the gifted amateur*. OUP Oxford, 2014.

[46] J. D. Jackson, *Classical Electrodynamics*. John Wiley & Sons, 1962.

[47] W. C. Chew, "Fields and waves: Lecture notes for ECE 350 at UIUC," *https://engineering.purdue.edu/wcchew/ece350.html*, 1990.

[48] C. M. Bender and S. A. Orszag, *Advanced mathematical methods for scientists and engineers I: Asymptotic methods and perturbation theory*. Springer Science & Business Media, 2013.

[49] J. D. Jackson, *Classical electrodynamics*. John Wiley & Sons, 1999.

[50] J. M. Crowley, *Fundamentals of applied electrostatics*. Krieger Publishing Company, 1986.

[51] C. Balanis, *Advanced Engineering Electromagnetics*. Hoboken, NJ, USA: Wiley, 2012.

[52] R. Courant and D. Hilbert, *Methods of Mathematical Physics, Volumes 1 and 2*. Interscience Publ., 1962.

[53] R. F. Harrington, *Time-harmonic electromagnetic fields*. McGraw-Hill, 1961.

[54] L. Esaki and R. Tsu, "Superlattice and negative differential conductivity in semiconductors," *IBM Journal of Research and Development*, vol. 14, no. 1, pp. 61–65, 1970.

[55] E. Kudeki and D. C. Munson, *Analog Signals and Systems*. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2009.

[56] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. Pearson Education, 2014.

[57] E. C. Jordan and K. G. Balmain, *Electromagnetic waves and radiating systems*. Prentice-Hall, 1968.

[58] G. Agarwal, D. Pattanayak, and E. Wolf, "Electromagnetic fields in spatially dispersive media," *Physical Review B*, vol. 10, no. 4, p. 1447, 1974.

[59] S. L. Chuang, *Physics of photonic devices*. John Wiley & Sons, 2012, vol. 80.

[60] B. E. Saleh and M. C. Teich, *Fundamentals of photonics*. John Wiley & Sons, 2019.

[61] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013, also 1959 to 1986.

[62] R. W. Boyd, *Nonlinear optics*. Elsevier, 2003.

[63] Y.-R. Shen, *The principles of nonlinear optics*. New York, Wiley-Interscience, 1984.

[64] N. Bloembergen, *Nonlinear optics*. World Scientific, 1996.

[65] P. C. Krause, O. Wasynczuk, and S. D. Sudhoff, *Analysis of electric machinery*. McGraw-Hill New York, 1986.

[66] A. E. Fitzgerald, C. Kingsley, S. D. Umans, and B. James, *Electric machinery*. McGraw-Hill New York, 2003, vol. 5.

[67] M. A. Brown and R. C. Semelka, *MRI.: Basic Principles and Applications*. John Wiley & Sons, 2011.

[68] C. A. Balanis, *Advanced engineering electromagnetics*. John Wiley & Sons, 1999, also 1989.

[69] Wikipedia, "Lorentz force," https://en.wikipedia.org/wiki/Lorentz_force/, accessed: 2019-09-06.

[70] R. P. Feynman, R. Leighton, and M. Sands, *Feynman Lectures on Physics, Volume III: Quantum Mechanics*. New York, NY, USA: Basic books, 1965.

[71] R. O. Dendy, *Plasma physics: an introductory course*. Cambridge University Press, 1995.

[72] Wikipedia, "Kennelly-Heaviside Layer," https://en.wikipedia.org/wiki/Kennelly-Heaviside_layer.

[73] H. Haken, *Quantum field theory of solids, an introduction*. North-Holland, 1976.

[74] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics, Vols. I, II, & III: The new millennium edition*. Basic books, 2011, also 1963, 2006, vol. 1,2,3.

[75] Wikipedia, "Spectral line shape," https://en.wikipedia.org/wiki/Spectral_line_shape.

[76] P. Sen and W. C. Chew, "The frequency dependent dielectric and conductivity response of sedimentary rocks," *Journal of microwave power*, vol. 18, no. 1, pp. 95–105, 1983.

[77] D. A. Miller, *Quantum Mechanics for Scientists and Engineers*. Cambridge, UK: Cambridge University Press, 2008.

[78] W. C. Chew, "Quantum mechanics made simple: Lecture notes for ECE 487 at UIUC," *http://wcchew.ece.illinois.edu/chew/course/QMAll20161206.pdf*, 2016.

[79] B. G. Streetman and S. Banerjee, *Solid state electronic devices*. Prentice Hall New Jersey, 2000, vol. 4.

[80] Smithsonian, "This 1600-year-old goblet shows that the romans were nanotechnology pioneers," https://www.smithsonianmag.com/history/this-1600-year-old-goblet-shows-that-the-romans-were-nanotechnology-pioneers-787224/, accessed: 2019-09-06.

[81] K. G. Budden, *Radio waves in the ionosphere*. Cambridge University Press, 2009.

[82] R. Fitzpatrick, *Plasma physics: an introduction*. CRC Press, 2014.

[83] G. Strang, *Introduction to linear algebra*. Wellesley-Cambridge Press Wellesley, MA, 1993, vol. 3.

[84] K. C. Yeh and C.-H. Liu, "Radio wave scintillations in the ionosphere," *Proceedings of the IEEE*, vol. 70, no. 4, pp. 324–360, 1982.

[85] J. Kraus, *Electromagnetics*. McGraw-Hill, 1984, also 1953, 1973, 1981.

[86] Wikipedia, "Circular polarization," https://en.wikipedia.org/wiki/Circular_polarization.

[87] Q. Zhan, "Cylindrical vector beams: from mathematical concepts to applications," *Advances in Optics and Photonics*, vol. 1, no. 1, pp. 1–57, 2009.

[88] Wikipedia, "Louis de Broglie," https://en.wikipedia.org/wiki/Louis_de_Broglie.

[89] H. Haus, *Electromagnetic Noise and Quantum Optical Measurements*, ser. Advanced Texts in Physics. Springer Berlin Heidelberg, 2000.

[90] W. C. Chew, "Lectures on theory of microwave and optical waveguides, for ECE 531 at UIUC," *https://engineering.purdue.edu/wcchew/course/tgwAll20160215.pdf*, 2016.

[91] L. Brillouin, *Wave propagation and group velocity*. Academic Press, 1960.

[92] F. B. Hildebrand, *Advanced calculus for applications*. Prentice-Hall, 1962.

[93] S. Schelkunoff, "Some equivalence theorems of electromagnetics and their application to radiation problems," *The Bell System Technical Journal*, vol. 15, no. 1, pp. 92–112, 1936.

[94] C.-T. Chen, *Linear system theory and design*. Oxford University Press, Inc., 1998.

[95] C. Kittel and P. McEuen, *Introduction to solid state physics*. Wiley New York, 1996, vol. 8.

[96] N. W. Ashcroft and N. D. Mermin, *Solid state physics*. Cengage Learning, 2022.

[97] S. H. Schot, "Eighty years of Sommerfeld's radiation condition," *Historia mathematica*, vol. 19, no. 4, pp. 385–401, 1992.

[98] D. Hinton, *Analects*. Counterpoint, 2014.

[99] H. Bible, "New Revised Standard Version (NRSV)," *Grand Rapids MI: Zondervan*, 1989.

[100] V. Rumsey, "Reaction concept in electromagnetic theory," *Physical Review*, vol. 94, no. 6, p. 1483, 1954.

[101] R. E. Collin, *Foundations for microwave engineering*. John Wiley & Sons, 2007, also 1966.

[102] W. J. Hoefer, "The transmission-line matrix method-theory and applications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 33, no. 10, pp. 882–893, 1985.

[103] A. Ruehli, G. Antonini, and L. Jiang, *Circuit oriented electromagnetic modeling using the PEEC techniques*. John Wiley & Sons, 2017.

[104] A. Ishimaru, *Electromagnetic wave propagation, radiation, and scattering from fundamentals to applications*. Wiley Online Library, 2017, also 1991.

[105] A. E. H. Love, "I. the integration of the equations of propagation of electric waves," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 197, no. 287-299, pp. 1–45, 1901.

[106] Wikipedia, "Christiaan Huygens," https://en.wikipedia.org/wiki/Christiaan_Huygens.

[107] ——, "George Green (mathematician)," https://en.wikipedia.org/wiki/George_Green_(mathematician).

[108] W. Chew, *Waves and Fields in Inhomogeneous Media, 378±381*. Van Nostrand, 1990.

[109] C.-T. Tai, *Dyadic Green's Functions in Electromagnetic Theory.* PA: International Textbook, Scranton, 1971.

[110] ——, *Dyadic Green functions in electromagnetic theory.* Institute of Electrical & Electronics Engineers (IEEE), 1994.

[111] W. Franz, "Zur formulierung des huygensschen prinzips," *Zeitschrift für Naturforschung A*, vol. 3, no. 8-11, pp. 500–506, 1948.

[112] J. C. Maxwell, *A treatise on electricity and magnetism.* Oxford: Clarendon Press, 1873, vol. 1.

[113] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables.* Courier Corporation, 1965, vol. 55.

[114] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products.* Academic press, 2014.

[115] W. C. Chew and J. Kong, "Microstrip capacitance for a circular disk through matched asymptotic expansions," *SIAM Journal on Applied Mathematics*, vol. 42, no. 2, pp. 302–317, 1982.

[116] J. D. Jackson, *Classical Electrodynamics, Third Edition.* John Wiley & Sons, 1974.

[117] J. F. Lee, "Finite element methods for modeling passive microwave devices," Ph.D. dissertation, Carnegie Mellon University.

[118] J.-F. Lee, "Analysis of passive microwave devices by using three-dimensional tangential vector finite elements," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 3, no. 4, pp. 235–246, 1990.

[119] L. Nagel and D. Pederson, "Simulation program with integrated circuit emphasis," in *Midwest Symposium on Circuit Theory*, 1973.

[120] T. M. Philip and M. J. Gilbert, "High-performance nanoscale topological inductor," in *2017 75th Annual Device Research Conference (DRC).* IEEE, 2017, pp. 1–2.

[121] R. Plonsey and R. E. Collin, *Principles and applications of electromagnetic fields.* McGraw-Hill, 1961.

[122] A. Wadhwa, A. L. Dal, and N. Malhotra, "Transmission media," https://www.slideshare.net/abhishekwadhwa786/transmission-media-9416228.

[123] P. H. Smith, "Transmission line calculator," *Electronics*, vol. 12, no. 1, pp. 29–31, 1939.

[124] D. M. Pozar, E. J. K. Knapp, and J. B. Mead, "ECE 584 microwave engineering laboratory notebook," http://www.ecs.umass.edu/ece/ece584/ECE584_lab_manual.pdf, 2004.

[125] Wikipedia, "Automated Network Analyzer," https://en.wikipedia.org/wiki/Network_analyzer_(electrical).

[126] J. Schutt-Aine, "Experimental-coaxial transmission line measurement using slotted line," http://emlab.uiuc.edu/ece451/ECE451Lab02.pdf.

[127] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, others, and J. M. Martinis, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510, 2019.

[128] Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan *et al.*, "Strong quantum computational advantage using a superconducting quantum processor," *arXiv preprint arXiv:2106.14734*, 2021.

[129] V. G. Veselago, "Electrodynamics of substances with simultaneously negative and," *Usp. Fiz. Nauk*, vol. 92, p. 517, 1967.

[130] J. B. Pendry, "Negative refraction makes a perfect lens," *Physical review letters*, vol. 85, no. 18, p. 3966, 2000.

[131] R. E. Collin, *Field theory of guided waves*. McGraw-Hill, 1960.

[132] D. M. Pozar, *Microwave engineering*. John wiley & sons, 2011.

[133] B. Clark, D. F. Allen, D. L. Best, S. D. Bonner, J. Jundt, M. G. Luling, M. O. Ross *et al.*, "Electromagnetic propagation logging while drilling: Theory and experiment," *SPE formation evaluation*, vol. 5, no. 03, pp. 263–271, 1990.

[134] B. Vaughn and D. Peroulis, "An updated applied formulation for the Goubau transmission line," *Journal of Applied Physics*, vol. 126, no. 19, p. 194902, 2019.

[135] Q. S. Liu, S. Sun, and W. C. Chew, "A potential-based integral equation method for low-frequency electromagnetic problems," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 3, pp. 1413–1426, 2018.

[136] Z. Ying, "Opening our eyes to spectacles of the past," https://www.shine.cn/feature/art-culture/1909061557/.

[137] Wikipedia, "Snell's law," https://en.wikipedia.org/wiki/Snell's_law.

[138] ——, "Graphical processing unit," https://en.wikipedia.org/wiki/Graphics_processing_unit.

[139] G. Tyras, *Radiation and propagation of electromagnetic waves*. Academic Press, 1969.

[140] L. Brekhovskikh, *Waves in layered media*. Academic Press, 1980.

[141] Scholarpedia, "Goos-hanchen effect," http://www.scholarpedia.org/article/Goos-Hanchen_effect.

[142] K. Kao and G. A. Hockham, "Dielectric-fibre surface waveguides for optical frequencies," in *Proceedings of the Institution of Electrical Engineers*, vol. 113, no. 7. IET, 1966, pp. 1151–1158.

[143] E. Glytsis, "Slab waveguide fundamentals," http://users.ntua.gr/eglytsis/IO/Slab_Waveguides_p.pdf, 2018.

[144] Wikipedia, "Optical fiber," https://en.wikipedia.org/wiki/Optical_fiber.

[145] Atlantic Cable, "1869 indo-european cable," https://atlantic-cable.com/Cables/1869IndoEur/index.htm.

[146] Wikipedia, "Submarine communications cable," https://en.wikipedia.org/wiki/Submarine_communications_cable.

[147] D. Brewster, "On the laws which regulate the polarisation of light by reflexion from transparent bodies," *Philosophical Transactions of the Royal Society of London*, vol. 105, pp. 125–159, 1815.

[148] Wikipedia, "Brewster's angle," https://en.wikipedia.org/wiki/Brewster's_angle.

[149] H. Raether, "Surface plasmons on smooth surfaces," in *Surface plasmons on smooth and rough surfaces and on gratings*. Springer, 1988, pp. 4–39.

[150] E. Kretschmann and H. Raether, "Radiative decay of non radiative surface plasmons excited by light," *Zeitschrift für Naturforschung A*, vol. 23, no. 12, pp. 2135–2136, 1968.

[151] Wikipedia, "Homomorphic Encryption," https://en.wikipedia.org/wiki/Homomorphic_encryption.

[152] K. Aki and P. G. Richards, *Quantitative seismology*, 2002.

[153] B. A. Auld, *Acoustic fields and waves in solids*. Ripol Classic, 1973.

[154] Wikipedia, "Surface plasmon," https://en.wikipedia.org/wiki/Surface_plasmon.

[155] A. Sommerfeld, *Über die Ausbreitung der Wellen in der drahtlosen Telegraphie*. Verlag der Königlich Bayerischen Akademie der Wissenschaften, 1909.

[156] Wikimedia, "Gaussian wave packet," https://commons.wikimedia.org/wiki/File:Gaussian_wave_packet.svg.

[157] W. C. Chew, "Lectures on theory of microwave and optical waveguides," *arXiv preprint arXiv:2107.09672*, 2021.

[158] Wikipedia, "Charles K. Kao," https://en.wikipedia.org/wiki/Charles_K._Kao.

[159] H. B. Callen and T. A. Welton, "Irreversibility and generalized noise," *Physical Review*, vol. 83, no. 1, p. 34, 1951.

[160] R. Kubo, "The fluctuation-dissipation theorem," *Reports on progress in physics*, vol. 29, no. 1, p. 255, 1966.

[161] C. Lee, S. Lee, and S. Chuang, "Plot of modal field distribution in rectangular and circular waveguides," *IEEE transactions on microwave theory and techniques*, vol. 33, no. 3, pp. 271–274, 1985.

[162] W. C. Chew, *Waves and Fields in Inhomogeneous Media*. IEEE Press, 1996.

[163] M. Abramowitz and I. A. Stegun, "Handbook of mathematical functions: with formulas, graphs, and mathematical tables," http://people.math.sfu.ca/~cbm/aands/index.htm.

[164] C. A. Balanis and E. Holzman, "Circular waveguides," *Encyclopedia of RF and Microwave Engineering*, 2005.

[165] W. C. Chew, W. Sha, and Q. I. Dai, "Green's dyadic, spectral function, local density of states, and fluctuation dissipation theorem," *arXiv preprint arXiv:1505.01586*, 2015.

[166] J. B. Johnson, "Thermal agitation of electricity in conductors," *Physical review*, vol. 32, no. 1, p. 97, 1928.

[167] H. Nyquist, "Thermal agitation of electric charge in conductors," *Physical review*, vol. 32, no. 1, p. 110, 1928.

[168] Wikipedia, "Very Large Array," https://en.wikipedia.org/wiki/Very_Large_Array.

[169] M. Al-Hakkak and Y. Lo, "Circular waveguides with anisotropic walls," *Electronics Letters*, vol. 6, no. 24, pp. 786–789, 1970.

[170] Wikipedia, "Horn Antenna," https://en.wikipedia.org/wiki/Horn_antenna.

[171] G. L. Matthaei, L. Young, and E. M. T. Jones, *Microwave filters, impedance-matching networks, and coupling structures*. Artech house, 1980.

[172] P. Silvester and P. Benedek, "Microstrip discontinuity capacitances for right-angle bends, t junctions, and crossings," *IEEE Transactions on Microwave Theory and Techniques*, vol. 21, no. 5, pp. 341–346, 1973.

[173] R. Garg and I. Bahl, "Microstrip discontinuities," *International Journal of Electronics Theoretical and Experimental*, vol. 45, no. 1, pp. 81–87, 1978.

[174] P. Smith and E. Turner, "A bistable fabry-perot resonator," *Applied Physics Letters*, vol. 30, no. 6, pp. 280–281, 1977.

[175] A. Yariv, *Optical electronics*. Saunders College Publ., 1991.

[176] Wikipedia, "Klystron," https://en.wikipedia.org/wiki/Klystron.

[177] ——, "Magnetron," https://en.wikipedia.org/wiki/Cavity_magnetron.

[178] ——, "Absorption Wavemeter," https://en.wikipedia.org/wiki/Absorption_wavemeter.

[179] G. Quaranta, G. Basset, O. J. Martin, and B. Gallinet, "Recent advances in resonant waveguide gratings," *Laser & Photonics Reviews*, vol. 12, no. 9, p. 1800017, 2018.

[180] G. Righini, Y. Dumeige, P. Feron, M. Ferrari, G. Nunzi Conti, D. Ristic, and S. Soria, "Whispering gallery mode microresonators: fundamentals and applications," *La Rivista del Nuovo Cimento*, vol. 34, pp. 435–488, 2011.

[181] G. Strang, *Linear algebra and its applications*. Academic Press, 1976.

[182] Y. P. Zhang and D. Liu, "Antenna-on-chip and antenna-in-package solutions to highly integrated millimeter-wave devices for wireless communications," *IEEE Transactions on Antennas and Propagation*, vol. 57, no. 10, pp. 2830–2841, 2009.

[183] Y. Zhang and J. Mao, "An overview of the development of antenna-in-package technology for highly integrated wireless devices," *Proceedings of the IEEE*, vol. 107, no. 11, pp. 2265–2280, 2019.

[184] W. C. Chew, "Vector potential electromagnetics with generalized gauge for inhomogeneous media: Formulation," *Progress In Electromagnetics Research*, vol. 149, pp. 69–84, 2014.

[185] W. Chew, "Computational electromagnetics: the physics of smooth versus oscillatory fields," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 362, no. 1816, pp. 579–602, 2004.

[186] L. Rayleigh, "X. on the electromagnetic theory of light," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 12, no. 73, pp. 81–101, 1881.

[187] G. B. Arfken, H.-J. Weber, and F. E. Harris, *Mathematical Methods for Physicists (Waltham, MA*. Elsevier, 2013.

[188] J. Zhu, T. E. Roth, D.-Y. Na, and W. C. Chew, "Generalized helmholtz decomposition for modal analysis of electromagnetic problems in inhomogeneous media," *IEEE Journal on Multiscale and Multiphysics Computational Techniques*, 2023.

[189] Wikipedia, "Guglielmo Marconi," https://en.wikipedia.org/wiki/Guglielmo_Marconi.

[190] B. P. Lathi and R. A. Green, *Linear systems and signals*. Oxford University Press New York, 2005, vol. 2.

[191] R. F. Harrington, "Matrix methods for field problems," *Proceedings of the IEEE*, vol. 55, no. 2, pp. 136–149, 1967.

[192] C. A. Balanis, *Antenna theory: analysis and design*. John Wiley & Sons, 2016.

[193] J.-M. Jin, *The finite element method in electromagnetics*. John Wiley & Sons, 2015.

[194] S. A. Schelkunoff and H. T. Friis, *Antennas: theory and practice*. Wiley New York, 1952, vol. 639.

[195] H. G. Schantz, "A brief history of UWB antennas," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 4, pp. 22–26, 2004.

[196] E. Kudeki, "Fields and Waves," http://remote2.ece.illinois.edu/~erhan/FieldsWaves/ECE350lectures.html.

[197] Wikipedia, "Antenna Aperture," https://en.wikipedia.org/wiki/Antenna_aperture.

[198] L. Josefsson and P. Persson, *Conformal array antenna theory and design*. John Wiley & Sons, 2006, vol. 29.

[199] R. J. Mailloux, *Phased array antenna handbook*. Artech House, 2017.

[200] J. G. Proakis, *Digital signal processing: principles algorithms and applications*. Pearson Education India, 2001.

[201] R. W. P. King, G. S. Smith, M. Owens, and T. Wu, "Antennas in matter: Fundamentals, theory, and applications," *NASA STI/Recon Technical Report A*, vol. 81, 1981.

[202] H. Yagi and S. Uda, "Projector of the sharpest beam of electric waves," *Proceedings of the Imperial Academy*, vol. 2, no. 2, pp. 49–52, 1926.

[203] Wikipedia, "Yagi-Uda Antenna," https://en.wikipedia.org/wiki/Yagi-Uda_antenna.

[204] ——, "Dipole Antenna," https://en.wikipedia.org/wiki/Dipole_antenna.

[205] ——, "Twin-Lead," https://en.wikipedia.org/wiki/Twin-lead.

[206] D. Dregely, R. Taubert, J. Dorfmüller, R. Vogelgesang, K. Kern, and H. Giessen, "3D optical Yagi–Uda nanoantenna array," *Nature communications*, vol. 2, no. 1, pp. 1–7, 2011.

[207] Antenna-theory.com, "Slot Antenna," http://www.antenna-theory.com/antennas/aperture/slot.php.

[208] Y. Lo, D. Solomon, and W. Richards, "Theory and experiment on microstrip antennas," *IEEE Transactions on Antennas and Propagation*, vol. 27, no. 2, pp. 137–145, 1979.

[209] A. D. Olver and P. J. Clarricoats, *Microwave horns and feeds*. IET, 1994, vol. 39.

[210] B. Thomas, "Design of corrugated conical horns," *IEEE Transactions on Antennas and Propagation*, vol. 26, no. 2, pp. 367–372, 1978.

[211] P. J. B. Clarricoats and A. D. Olver, *Corrugated horns for microwave antennas*. IET, 1984, no. 18.

[212] Y. T. Lo and S. Lee, *Antenna Handbook: Volume III Applications*. Springer Science & Business Media, 2012.

[213] P. Gibson, "The Vivaldi aerial," in *1979 9th European Microwave Conference*. IEEE, 1979, pp. 101–105.

[214] Wikipedia, "Vivaldi Antenna," https://en.wikipedia.org/wiki/Vivaldi_antenna.

[215] W. H. Weedon, W. C. Chew, and P. E. Mayes, "A step-frequency radar imaging system for microwave nondestructive evaluation," *Progress In Electromagnetics Research*, vol. 28, pp. 121–146, 2000.

[216] Wikipedia, "Cassegrain Antenna," https://en.wikipedia.org/wiki/Cassegrain_antenna.

[217] ——, "Cassegrain Reflector," https://en.wikipedia.org/wiki/Cassegrain_reflector.

[218] W. A. Imbriale, S. S. Gao, and L. Boccia, *Space antenna handbook*. John Wiley & Sons, 2012.

[219] J. A. Encinar, "Design of two-layer printed reflectarrays using patches of variable size," *IEEE Transactions on Antennas and Propagation*, vol. 49, no. 10, pp. 1403–1410, 2001.

[220] D.-C. Chang and M.-C. Huang, "Microstrip reflectarray antenna with offset feed," *Electronics Letters*, vol. 28, no. 16, pp. 1489–1491, 1992.

[221] G. Minatti, M. Faenzi, E. Martini, F. Caminita, P. De Vita, D. González-Ovejero, M. Sabbadini, and S. Maci, "Modulated metasurface antennas for space: Synthesis, analysis and realizations," *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 4, pp. 1288–1300, 2014.

[222] X. Gao, X. Han, W.-P. Cao, H. O. Li, H. F. Ma, and T. J. Cui, "Ultrawideband and high-efficiency linear polarization converter based on double v-shaped metasurface," *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 8, pp. 3522–3530, 2015.

[223] D. De Schweinitz and T. L. Frey Jr, "Artificial dielectric lens antenna," Nov. 13 2001, US Patent 6,317,092.

[224] K.-L. Wong, "Planar antennas for wireless communications," *Microwave Journal*, vol. 46, no. 10, pp. 144–145, 2003.

[225] H. Nakano, M. Yamazaki, and J. Yamauchi, "Electromagnetically coupled curl antenna," *Electronics Letters*, vol. 33, no. 12, pp. 1003–1004, 1997.

[226] K. Lee, K. Luk, K.-F. Tong, S. Shum, T. Huynh, and R. Lee, "Experimental and simulation studies of the coaxially fed U-slot rectangular patch antenna," *IEE Proceedings-Microwaves, Antennas and Propagation*, vol. 144, no. 5, pp. 354–358, 1997.

[227] K. Luk, C. Mak, Y. Chow, and K. Lee, "Broadband microstrip patch antenna," *Electronics letters*, vol. 34, no. 15, pp. 1442–1443, 1998.

[228] M. Bolic, D. Simplot-Ryl, and I. Stojmenovic, *RFID systems: research trends and challenges*. John Wiley & Sons, 2010.

[229] D. M. Dobkin, S. M. Weigand, and N. Iyer, "Segmented magnetic antennas for near-field UHF RFID," *Microwave Journal*, vol. 50, no. 6, p. 96, 2007.

[230] Z. N. Chen, X. Qing, and H. L. Chung, "A universal UHF RFID reader antenna," *IEEE transactions on microwave theory and techniques*, vol. 57, no. 5, pp. 1275–1282, 2009.

[231] Wikipedia, "Faraday Cage," https://en.wikipedia.org/wiki/Faraday_cage.

[232] J. A. Stratton, *Electromagnetic Theory*. McGraw-Hill Book Company, Inc., 1941.

[233] W. Meissner and R. Ochsenfeld, "Ein neuer effekt bei eintritt der supraleitfähigkeit," *Naturwissenschaften*, vol. 21, no. 44, pp. 787–788, 1933.

[234] Wikipedia, "Superconductivity," https://en.wikipedia.org/wiki/Superconductivity.

[235] D. Sievenpiper, L. Zhang, R. F. Broas, N. G. Alexopolous, and E. Yablonovitch, "High-impedance electromagnetic surfaces with a forbidden frequency band," *IEEE Transactions on Microwave Theory and techniques*, vol. 47, no. 11, pp. 2059–2074, 1999.

[236] Wikipedia, "Snell's law," https://en.wikipedia.org/wiki/Snell's_law.

[237] H. Lamb, "On sommerfeld's diffraction problem; and on reflection by a parabolic mirror," *Proceedings of the London Mathematical Society*, vol. 2, no. 1, pp. 190–203, 1907.

[238] W. J. Smith, *Modern optical engineering*. McGraw-Hill New York, 1966, vol. 3.

[239] D. C. O'Shea, T. J. Suleski, A. D. Kathman, and D. W. Prather, *Diffractive optics: design, fabrication, and test*. SPIE Press Bellingham, WA, 2004, vol. 62.

[240] J. B. Keller and H. B. Keller, "Determination of reflected and transmitted fields by geometrical optics," *JOSA*, vol. 40, no. 1, pp. 48–52, 1950.

[241] G. A. Deschamps, "Ray techniques in electromagnetics," *Proceedings of the IEEE*, vol. 60, no. 9, pp. 1022–1035, 1972.

[242] R. G. Kouyoumjian and P. H. Pathak, "A uniform geometrical theory of diffraction for an edge in a perfectly conducting surface," *Proceedings of the IEEE*, vol. 62, no. 11, pp. 1448–1461, 1974.

[243] R. Kouyoumjian, "The geometrical theory of diffraction and its application," in *Numerical and Asymptotic Techniques in Electromagnetics*. Springer, 1975, pp. 165–215.

[244] S.-W. Lee and G. Deschamps, "A uniform asymptotic theory of electromagnetic diffraction by a curved wedge," *IEEE Transactions on Antennas and Propagation*, vol. 24, no. 1, pp. 25–34, 1976.

[245] Wikipedia, "Fermat's principle," https://en.wikipedia.org/wiki/Fermat's_principle.

[246] N. Yu, P. Genevet, M. A. Kats, F. Aieta, J.-P. Tetienne, F. Capasso, and Z. Gaburro, "Light propagation with phase discontinuities: generalized laws of reflection and refraction," *Science*, vol. 334, no. 6054, pp. 333–337, 2011.

[247] X. Ni, N. K. Emani, A. V. Kildishev, A. Boltasseva, and V. M. Shalaev, "Broadband light bending with plasmonic nanoantennas," *Science*, vol. 335, no. 6067, pp. 427–427, 2012.

[248] A. Sommerfeld, *Partial differential equations in physics*. Academic Press, 1949, vol. 1.

[249] R. Haberman, *Elementary applied partial differential equations*. Prentice Hall Englewood Cliffs, NJ, 1983, vol. 987.

[250] G. A. Deschamps, "Gaussian beam as a bundle of complex rays," *Electronics letters*, vol. 7, no. 23, pp. 684–685, 1971.

[251] J. Enderlein and F. Pampaloni, "Unified operator approach for deriving Hermite–Gaussian and Laguerre–Gaussian laser modes," *JOSA A*, vol. 21, no. 8, pp. 1553–1558, 2004.

[252] D. L. Andrews, *Structured light and its applications: An introduction to phase-structured beams and nanoscale optical forces*. Academic Press, 2011.

[253] J. W. Strutt, "Xv. on the light from the sky, its polarization and colour," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 41, no. 271, pp. 107–120, 1871.

[254] S. Sun, Y. G. Liu, W. C. Chew, and Z. Ma, "Calderón multiplicative preconditioned efie with perturbation method," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 1, pp. 247–255, 2012.

[255] G. Mie, "Beiträge zur optik trüber medien, speziell kolloidaler metallösungen," *Annalen der physik*, vol. 330, no. 3, pp. 377–445, 1908.

[256] Wikipedia, "Mie scattering," https://en.wikipedia.org/wiki/Mie_scattering.

[257] K. Sarabandi, *Foundations of Applied Electromagnetics*. Michigan Publishing, 2023.

[258] W.C. Chew, I. Aksun, J.H. Lin, C.C. Lu, G. Otto, Y.M. Wang, R. Wagner, W.H. Weedon, "Solution manual to waves and fields in inhomogeneous media," https://engineering.purdue.edu/wcchew/solution_manual_WFIM.pdf, 1993.

[259] L. B. Felsen and N. Marcuvitz, *Radiation and scattering of waves*. John Wiley & Sons, 1994, also 1973, vol. 31.

[260] P. P. Ewald, "Die berechnung optischer und elektrostatischer gitterpotentiale," *Annalen der physik*, vol. 369, no. 3, pp. 253–287, 1921.

[261] E. Whitaker and G. Watson, *A Course of Modern Analysis*. Cambridge Mathematical Library, 1927.

[262] J. Kong, "Electromagnetic fields due to dipole antennas over stratified anisotropic media," *Geophysics*, vol. 37, no. 6, pp. 985–996, 1972.

[263] W. Chew, "A quick way to approximate a Sommerfeld-Weyl-type integral (antenna far-field radiation)r," *IEEE Transactions on Antennas and Propagation*, vol. 36, no. 11, pp. 1654–1657, 1988.

[264] Wikipedia, "FLOPS," https://en.wikipedia.org/wiki/FLOPS.

[265] Techopedia, "Fastest Computer," https://www.techopedia.com/what-is-the-fastest-supercomputer-in-the-world.

[266] Wikipedia, "Computer," https://en.wikipedia.org/wiki/computer.

[267] H. Gan, "Numerical Green's function in surface integral equation method and hydrodynamic model for solar cell analysis," Ph.D. dissertation, University of Illinois, 2019.

[268] W. C. H. McLean, *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, 2000.

[269] G. C. Hsiao and W. L. Wendland, *Boundary integral equations*. Springer, 2008.

[270] W. C. Chew, M. S. Tong, and B. Hu, "Integral equation methods for electromagnetic and elastic waves," *Synthesis Lectures on Computational Electromagnetics*, vol. 3, no. 1, pp. 1–241, 2008.

[271] P. K. Banerjee and R. Butterfield, *Boundary element methods in engineering science*. McGraw-Hill London, 1981, vol. 17.

[272] O. C. Zienkiewicz, R. L. Taylor, P. Nithiarasu, and J. Zhu, *The finite element method*. McGraw-Hill London, 1977, vol. 3.

[273] J.-F. Lee, R. Lee, and A. Cangellaris, "Time-domain finite-element methods," *IEEE Transactions on Antennas and Propagation*, vol. 45, no. 3, pp. 430–442, 1997.

[274] J. L. Volakis, A. Chatterjee, and L. C. Kempel, *Finite element method electromagnetics: antennas, microwave circuits, and scattering applications*. John Wiley & Sons, 1998, vol. 6.

[275] J.-C. Nédélec, "Mixed finite elements in r 3," *Numerische Mathematik*, vol. 35, pp. 315–341, 1980.

[276] S. Rao, D. Wilton, and A. Glisson, "Electromagnetic scattering by surfaces of arbitrary shape," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 3, pp. 409–418, 1982.

[277] Cramer and Gabriel, *Introduction a l'analyse des lignes courbes algebriques par Gabriel Cramer...* chez les freres Cramer & Cl. Philibert, 1750.

[278] J. A. Schouten, *Tensor analysis for physicists*. Courier Corporation, 1989.

[279] A. C. Polycarpou, "Introduction to the finite element method in electromagnetics," *Synthesis Lectures on Computational Electromagnetics*, vol. 1, no. 1, pp. 1–126, 2005.

[280] J. P. A. Bastos and N. Sadowski, *Electromagnetic modeling by finite element methods*. CRC press, 2003.

[281] Ö. Özgün and M. Kuzuoğlu, *MATLAB-based Finite Element Programming in Electromagnetic Modeling*. CRC Press, 2018.

[282] R. Coifman, V. Rokhlin, and S. Wandzura, "The fast multipole method for the wave equation: A pedestrian prescription," *IEEE Antennas and Propagation magazine*, vol. 35, no. 3, pp. 7–12, 1993.

[283] J. Song, C.-C. Lu, and W. C. Chew, "Multilevel fast multipole algorithm for electromagnetic scattering by large complex objects," *IEEE Transactions on Antennas and Propagation*, vol. 45, no. 10, pp. 1488–1493, 1997.

[284] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge University Press, 2007.

[285] O. Axelsson and V. A. Barker, *Finite element solution of boundary value problems: theory and computation*. SIAM, 2001.

[286] R. Wait, A. R. Mitchell, and A. R. Mitchell, *Finite element analysis and applications*. John Wiley & Sons Incorporated, 1985.

[287] K. Yee, "Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media," *IEEE Transactions on Antennas and Propagation*, vol. 14, no. 3, pp. 302–307, 1966.

[288] A. Taflove, "Review of the formulation and applications of the finite-difference time-domain method for numerical modeling of electromagnetic wave interactions with arbitrary structures," *Wave Motion*, vol. 10, no. 6, pp. 547–582, 1988.

[289] A. Taflove and S. C. Hagness, *Computational electrodynamics: the finite-difference time-domain method*. Artech house, 2005, also 1995.

[290] W. Yu, R. Mittra, T. Su, Y. Liu, and X. Yang, *Parallel finite-difference time-domain method*. Artech House Norwood, 2006.

[291] D. Potter, "Computational physics," 1973.

[292] W. F. Ames, *Numerical methods for partial differential equations*. Academic press, 2014, also 1977.

[293] K. W. Morton, *Revival: Numerical Solution Of Convection-Diffusion Problems (1996)*. CRC Press, 2019.

[294] F. B. Hildebrand, *Introduction to numerical analysis*. Courier Corporation, 1987.

[295] W. C. Chew, "Electromagnetic theory on a lattice," *Journal of Applied Physics*, vol. 75, no. 10, pp. 4843–4850, 1994.

[296] J. v. Neumann, *Mathematische Grundlagen der Quantenmechanik, Berlın*. Springer, New York, Dover Publications, 1943.

[297] R. Courant, K. Friedrichs, and H. Lewy, "Über die partiellen differenzengleichungen der mathematischen physik," *Mathematische annalen*, vol. 100, no. 1, pp. 32–74, 1928.

[298] T. Weiland, "A discretization model for the solution of maxwell's equations for six-component fields," *Archiv Elektronik und Uebertragungstechnik*, vol. 31, pp. 116–120, 1977.

[299] M. Clemens and T. Weiland, "Discrete electromagnetism with the finite integration technique," *Progress In Electromagnetics Research*, vol. 32, pp. 65–87, 2001.

[300] M. Desbrun, A. N. Hirani, M. Leok, and J. E. Marsden, "Discrete exterior calculus," *arXiv preprint math/0508341*, 2005.

[301] J.-P. Berenger, "A perfectly matched layer for the absorption of electromagnetic waves," *Journal of computational physics*, vol. 114, no. 2, pp. 185–200, 1994.

[302] W. C. Chew and W. H. Weedon, "A 3D perfectly matched medium from modified maxwell's equations with stretched coordinates," *Microwave and optical technology letters*, vol. 7, no. 13, pp. 599–604, 1994.

[303] W. C. Chew, J. Jin, and E. Michielssen, "Complex coordinate system as a generalized absorbing boundary condition," in *IEEE Antennas and Propagation Society International Symposium 1997. Digest*, vol. 3. IEEE, 1997, pp. 2060–2063.

[304] A. Zee, *Quantum field theory in a nutshell*. Princeton university press, 2010, vol. 7.

[305] A. Aspect, P. Grangier, and G. Roger, "Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: a new violation of Bell's inequalities," *Physical Review Letters*, vol. 49, no. 2, p. 91, 1982.

[306] J. F. Clauser and A. Shimony, "Bell's theorem. experimental tests and implications," *Reports on Progress in Physics*, vol. 41, no. 12, p. 1881, 1978.

[307] D. M. Greenberger, M. A. Horne, A. Shimony, and A. Zeilinger, "Bell's theorem without inequalities," *American Journal of Physics*, vol. 58, no. 12, pp. 1131–1143, 1990.

[308] Wikipedia, "Double-slit experiment," https://en.wikipedia.org/wiki/Double-slit_experiment.

[309] B. Bapat, "Newton's rings," http://www.iiserpune.ac.in/~bhasbapat/phy221_files/NewtonsRing.pdf.

[310] Shmoop.Com, "Young's double-slit," https://www.shmoop.com/optics/young-double-slit.html.

[311] Wikipedia, "John Dalton," https://en.wikipedia.org/wiki/John_Dalton.

[312] ——, "Max Planck," https://en.wikipedia.org/wiki/Max_Planck.

[313] ——, "Photoelectric effect," https://en.wikipedia.org/wiki/Photoelectric_effect.

[314] ——, "Newton's laws of motion," https://en.wikipedia.org/wiki/Newton's_laws_of_motion.

[315] ——, "William Rowan Hamilton," https://en.wikipedia.org/wiki/William_Rowan_Hamilton.

[316] W. C. Chew, A. Y. Liu, C. Salazar-Lazaro, D. Na, and W. E. I. Sha, "Hamilton equation, commutator, and energy conservation," *Quantum Report*, pp. 295–303, Dec 2019.

[317] C. J. Ryu, E. Kudeki, D.-Y. Na, T. E. Roth, and W. C. Chew, "Fourier transform, dirac commutator, energy conservation, and correspondence principle for electrical engineers," *IEEE Journal on Multiscale and Multiphysics Computational Techniques*, vol. 7, pp. 69–83, 2022.

[318] M. Kira and S. W. Koch, *Semiconductor quantum optics*. Cambridge University Press, 2011.

[319] Wikipedia, "Gaussian beam," https://en.wikipedia.org/wiki/Gaussian_beam.

[320] ——, "Quantum harmonic oscillator," https://en.wikipedia.org/wiki/Quantum_harmonic_oscillator.

[321] W. C. Chew, D.-Y. Na, P. Bermel, T. E. Roth, C. J. Ryu, and E. Kudeki, "Quantum maxwell's equations made simple: Employing scalar and vector potential formulation," *IEEE Antennas and Propagation Magazine*, vol. 63, no. 1, pp. 14–26, 2020.

[322] W. H. Louisell, *Quantum statistical properties of radiation*. Wiley New York, 1973, vol. 7.

[323] J. S. Bell, "On the Einstein Podolsky Rosen paradox," *Physics Physique Fizika*, vol. 1, no. 3, p. 195, 1964.

[324] A. Aspect, J. Dalibard, and G. Roger, "Experimental test of Bell's inequalities using time-varying analyzers," *Physical Review Letters*, vol. 49, no. 25, p. 1804, 1982.

[325] Wikipedia, "Roy J. Glauber," https://en.wikipedia.org/wiki/Roy_J._Glauber.

[326] ——, "E.C. George Sudarshan," https://en.wikipedia.org/wiki/E._C._George_Sudarshan.

[327] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, "Theory of superconductivity," *Physical review*, vol. 108, no. 5, p. 1175, 1957.

[328] Techopedia, "Optical Fiber Loss," https://www.fiberoptics4sale.com/blogs/archive-posts/95049798-calculating-fiber-loss-and-distance-estimates.

[329] Scientific American, "Chinese Quantum Teleportation," https://www.scientificamerican.com/article/china-reaches-new-milestone-in-space-based-quantum-communications/.

[330] H.-S. Zhong, Y.-H. Deng, J. Qin, H. Wang, M.-C. Chen, L.-C. Peng, Y.-H. Luo, D. Wu, S.-Q. Gong, H. Su, Y. Hu, P. Hu, X.-Y. Yang, W.-J. Zhang, H. Li, Y. Li, X. Jiang, L. Gan, G. Yang, L. You, Z. Wang, L. Li, N.-L. Liu, J. J. Renema, C.-Y. Lu, and J.-W. Pan, "Phase-programmable gaussian boson sampling using stimulated squeezed light," *Phys. Rev. Lett.*, vol. 127, p. 180502, Oct 2021. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.127.180502

[331] C. Gerry and P. Knight, *Introductory Quantum Optics*. Cambridge, UK: Cambridge University Press, 2004.

[332] M. Fox, *Quantum optics: an introduction*. OUP Oxford, 2006, vol. 15.

[333] C. Gerry and P. L. Knight, *Introductory quantum optics*. Cambridge University Press, 2005.

[334] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen, "From potential theory to matrix iterations in six steps," *SIAM Review*, vol. 40, no. 3, pp. 547–578, 1998.

[335] Wikipedia, "Coherent state," https://en.wikipedia.org/wiki/Coherent_state.

[336] W. C. Chew, A. Y. Liu, C. Salazar-Lazaro, and W. E. I. Sha, "Quantum electromagnetics: A new look—part i," *J. Multiscale and Multiphys. Comput. Techn.*, vol. 1, pp. 73–84, 2016.

[337] ——, "Quantum electromagnetics: A new look—part ii," *J. Multiscale and Multiphys. Comput. Techn.*, vol. 1, pp. 85–97, 2016.

[338] E. Merzbacher, *Quantum Mechanics*. John Wiley & Sons, 1998.

[339] C. Cohen-Tannoudji, J. Dupont-Roc, and G. Grynberg, *Photons and atoms-introduction to quantum electrodynamics*, 1997.

[340] C. J. Ryu, D.-Y. Na, and W. C. Chew, "Matrix product states and numerical mode decomposition for the analysis of gauge-invariant cavity quantum electrodynamics," *Physical Review A*, vol. 107, no. 6, p. 063707, 2023.

[341] L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press, 1995.

[342] R. Loudon, *The Quantum Theory of Light*. Oxford, UK: OUP Oxford, 2000.

[343] D.-Y. Na, J. Zhu, W. C. Chew, and F. L. Teixeira, "Quantum information preserving computational electromagnetics," *Physical Review A*, vol. 102, no. 1, p. 013711, 2020.

Index